华为职业认证通过者权益

通过**任一项**华为职业认证，您即可在华为在线学习网站(*http://learning.huawei.com/cn*) 享有如下特权：

- 1、**华为E-learning 课程学习**
  - **内容**：**所有华为职业认证E-Learning课程**，扩展您在其他技术领域的技术知识
  - **方式**：请提交您的“华为账号”和注册账号的“email地址”到 *Learning@huawei.com 申请权限。*
- 2、**华为培训教材下载**
  - **内容：华为职业认证培训教材+华为产品技术培训教材，**覆盖企业网络、存储、安全等诸多领域
  - **方式**：登录*华为在线学习网站*，进入“*华为培训->面授培训*”，在具体课程页面即可下载教材。
- 3、**华为在线公开课(LVC)优先参与**
  - **内容**：企业网络、UC&C、安全、存储等诸多领域的职业认证课程，华为讲师授课，开班人数有限
  - **方式**：开班计划及参与方式请详见LVC排期：
    http://support.huawei.com/learning/NavigationAction!createNavi#navi[id]=_16
- 4、**学习工具 eNSP**
  - *eNSP (Enterprise Network Simulation Platform)*，是由华为提供的免费的、可扩展的、图形化网络仿真工具。主要对企业网路由器和交换机进行硬件模拟，完美呈现真实设备实景；同时也支持大型网络模拟，让大家在没有真实设备的情况下也能够进行实验测试。
- 另外，华为建立了知识分享平台 华为认证论坛。**您可以在线与华为技术专家交流技术，与其他考生分享考试经验，一起学习华为产品技术。**（http://support.huawei.com/ecommunity/bbs/list_2247.html）



华为认证网络工程师系列教程 – HCNA

华为网络存储工程师认证

Huawei Certified Network Associate-Storage



华为技术有限公司

华为认证网络工程师系列教程 – HCNA

华为网络存储工程师认证

第2.0版本

华为认证系统介绍

依托华为公司雄厚的技术实力和专业的培训体系，华为认证考虑到不同客户对 ICT 技术不同层次的需求，致力于为客户提供实战性、专业化的技术认证。 根据 ICT 技术的特点和客户不同层次的需求，华为认证为客户提供面向十三个方向的四级认证体系。

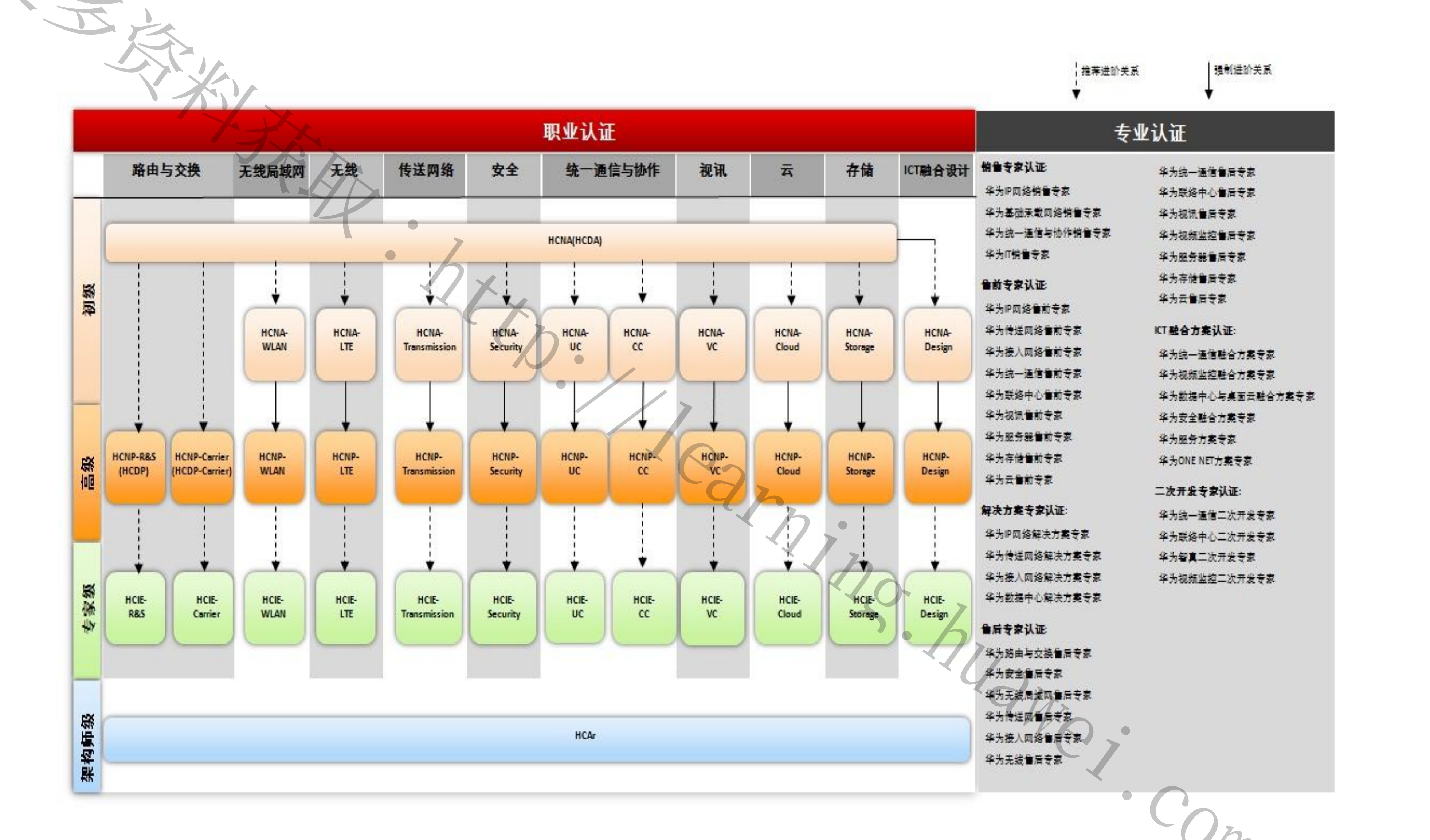
HCNA-Storage（Huawei Certified Network Associate-Storage，华为网络存储工程师认证）主要面向网络存储维护工程师，以及准备参加 HCNA-Storage 认证考试的人员；希望掌握 SAN 存储系统与网络基本原理和华为 SAN 存储阵系统管理、部署与维护能力的人员。HCNA 认证在内容上涵盖存储基础知识、RAID 技术与应用、存储网络技术与应用、华为存储产品与解决方案、存储系统管理和基本配置、存储主机连接与多路径配置、SAN 网络与存储系统日常维护。

HCNP-Storage (Huawei Certified Network Professional-Storage，华为认证网络存储资深工程师)主要面向企业级网络存储维护工程师、专家工程师以及希望系统深入 SAN 存储、统一存储、数据保护技术与部署的人员。

HCNP-Storage 包括 CUSN（Constructing Unifying Storage Network 构建统一存储网络）、CBDS(Constructing Big Data Solution 构建大数据解决方案)、CDPS（Constructing Data Protection System 构建数据保护系统）三个部分。内容上涵盖 SAN、NAS 统一存储原理、架构和组件，存储数据处理与通信协议（SCSI、FC、iSCSI）原理及应用，存储系统数据可靠性与业务连续性保障技术存储与主流 OS 平台连接与应用，存储网络冗余技术及应用，SAN、集群 NAS 网络规划与方案部署，虚拟快照、LUNCopy、复制的原理和部署，网络存储虚拟化技术及应用，存储虚拟化系统原理、部署和异构资源管理，备份网络及备份恢复技术及应用，华为数据保护方案构建、部署与管理，华为数据容灾方案及典型应用场景，华为存储系统、网络、方案的故障诊断与处理方法。

HCIE-Storage（Huawei Certified Internetwork Expert--Storage，华为认证互联网络存储专家）旨在培养能够熟练掌握各种存储技术；精通华为存储产品的维护、诊断和故障排除。

华为认证协助您打开行业之窗，开启改变之门，屹立在 ICT 世界的潮头浪尖！


推荐进阶关系
强制进阶关系
职业认证
专业认证
路由与交换
无线局域网
无线
传送网络
安全
统一通信与协作
视讯
云
存储
ICT融合设计
初级
高级
专家级
架构师级
HCNA(HCDA)
HCNA-WLAN
HCNA-LTE
HCNA-Transmission
HCNA-Security
HCNA-UC
HCNA-CC
HCNA-VC
HCNA-Cloud
HCNA-Storage
HCNA-Design
HCNP-R&S (HCDP)
HCNP-Carrier (HCDP-Carrier)
HCNP-WLAN
HCNP-LTE
HCNP-Transmission
HCNP-Security
HCNP-UC
HCNP-CC
HCNP-VC
HCNP-Cloud
HCNP-Storage
HCNP-Design
HCIE-R&S
HCIE-Carrier
HCIE-WLAN
HCIE-LTE
HCIE-Transmission
HCIE-Security
HCIE-UC
HCIE-CC
HCIE-VC
HCIE-Cloud
HCIE-Storage
HCIE-Design
HCAr
销售专家认证:
华为IP网络销售专家
华为基础承载网络销售专家
华为统一通信与协作销售专家
华为IT销售专家
售前专家认证:
华为IP网络售前专家
华为传送网络售前专家
华为接入网络售前专家
华为统一通信售前专家
华为联络中心售前专家
华为视讯售前专家
华为服务器售前专家
华为存储售前专家
华为云售前专家
解决方案专家认证:
华为IP网络解决方案专家
华为传送网络解决方案专家
华为接入网络解决方案专家
华为数据中心解决方案专家
售后专家认证:
华为路由与交换售后专家
华为安全售后专家
华为无线局域网售后专家
华为传送网售后专家
华为接入网络售后专家
华为无线售后专家
华为统一通信售后专家
华为联络中心售后专家
华为视讯售后专家
华为视频监控售后专家
华为服务器售后专家
华为存储售后专家
华为云售后专家
ICT融合方案认证:
华为统一通信融合方案专家
华为视频监控融合方案专家
华为数据中心与桌面云融合方案专家
华为安全融合方案专家
华为服务方案专家
华为ONE NET方案专家
二次开发专家认证:
华为统一通信二次开发专家
华为联络中心二次开发专家
华为智真二次开发专家
华为视频监控二次开发专家

前言

简介

本书为 HCNA-Storage 认证培训教程，适用于华为认证网络存储工程师以及准备参加 HCNA-Storage 考试的学员或者希望系统掌握华为网络存储产品与技术的学员。

内容描述

本书共包含 10 个 Module，系统地介绍了华为存储技术及其应用和常见的故障处理方法。

Module1 介绍了 IT 基础设施与数据中心，存储应用环境，机械硬盘以及 SSD 硬盘结构及实现原理，主机应用环境。

Module2 介绍了 RAID 分类及原理，RAID2.0+技术。

Module3 介绍存储阵列系统的组成，通用技术以及存储阵列的应用。

Module4 介绍 DAS 存储基础知识，SCSI 协议，SAN 存储基础知识，FC 连接和协议封装，SAN 存储的应用。

Module5 介绍 IP SAN 产生与发展的背景，IP SAN 组成和组网连接，iSCSI 协议，FC 协议与 TCP 协议融合，华 IP SAN 存储的实现与应用。

Module6 介绍 ISM 功能及安装方法，存储初始化配置流程，存储端基本配置，主机端基本配置，存储运维管理方法。

Module7 介绍 NAS 的产生与发展，NAS 系统组成与部件，NAS 文件共享协议 CIFS,NFS,NAS 文件系统 IO 与性能，SAN 与 NAS 比较，华为实现与应用。

Module8 介绍大数据的基本概念，对象存储技术，大数据处理技术，华为大数据实践。

Module9 介绍备份概念及拜年结构，备份技术，备份策略，华为备份实现与应用，容灾。

Module10 介绍云计算的概念与背景，部署与商业模式，核心技术与价值，解决方案。

本书将引导学员完成所有 Module，学完本课程之后将具备华为 SAN 存储产品基本规划和安装部署能力，以胜任网络存储工程师或系统管理员的工作岗位。

读者必备知识背景

本课程为华为网络存储认证基础课程，阅读本书的读者应首先具备计算机基础知识、硬件架构和软件基本原理、软件的基本操作。

本书常用图标



光纤交换机



以太网交换机



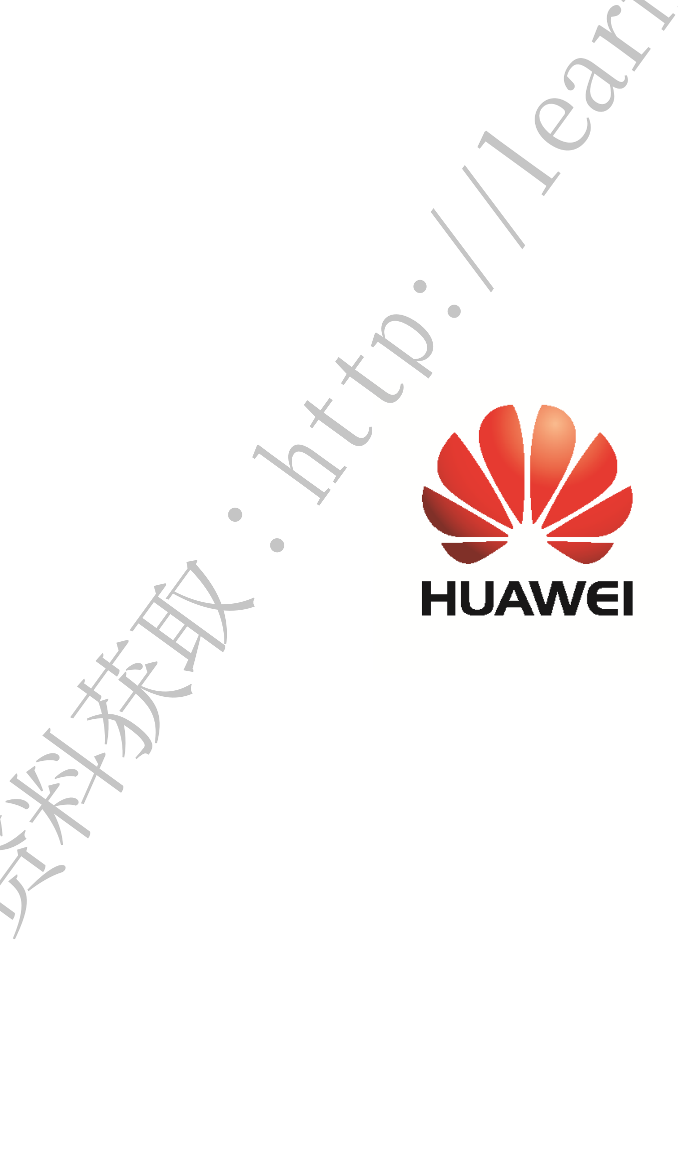
存储系统



主机

目 录

HC1109101 存储与应用环境




HC1109101
存储与应用环境
www.huawei.com
Copyright © 2013 Huawei Technologies Co., Ltd. All rights reserved.
HUAWEI



目标

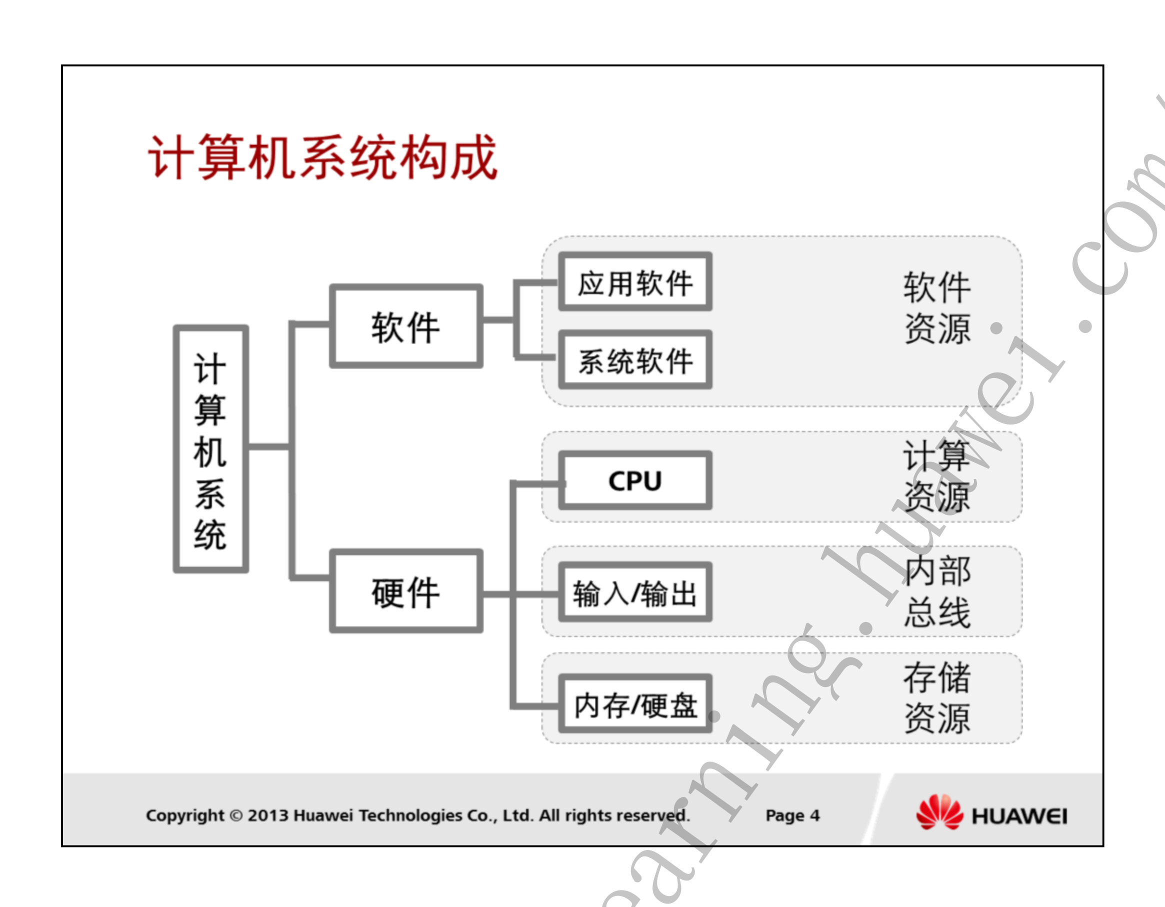
- 学习完本课程后，您将能够：
  - 了解IT基础设施与数据中心
  - 了解存储的应用环境
  - 了解机械硬盘以及SSD硬盘结构及实现原理
  - 了解主机应用环境





目录





企业IT系统离不开计算机/服务器，因此，在介绍IT基础设施之前，我们先回顾一下计算机系统的构成。

计算机系统的构成可以划分为四个层面，即：计算资源、内部总线、存储资源和软件资源。计算资源主要由计算的核心硬件CPU来实现。大家知道，CPU运算速度越快，计算机的计算能力也越强。数据临时存储由内存来完成，而长久存储则由内置硬盘、光盘、软盘等来实现。计算机内部的数据通讯则由其内部总线（BUS）来完成，包括数据的输入、输出等。一个完整的计算系统包括硬件与之相匹配的软件。在系统软件的基础上，为实现特定的业务功能应用，针对应用软件的开发也必不可少。

由上可以看出，在独立的计算系统中，数据计算、存储通常依靠CPU和内置的的存储设备来实现，在计算要求越来越快、计算规模越来越大、需存储数据量也越来越大时很容易出现瓶颈。

IT系统构成

软件资源　集群系统、分布式计算、应用系统、……

计算资源　Server1　Server2　Server3　……　Server(n)　集群

网络资源　Ethernet Switch　FC Switch　IP/FC

存储资源　IP SAN　FC SAN　磁带库　外部存储

　Page 5　HUAWEI

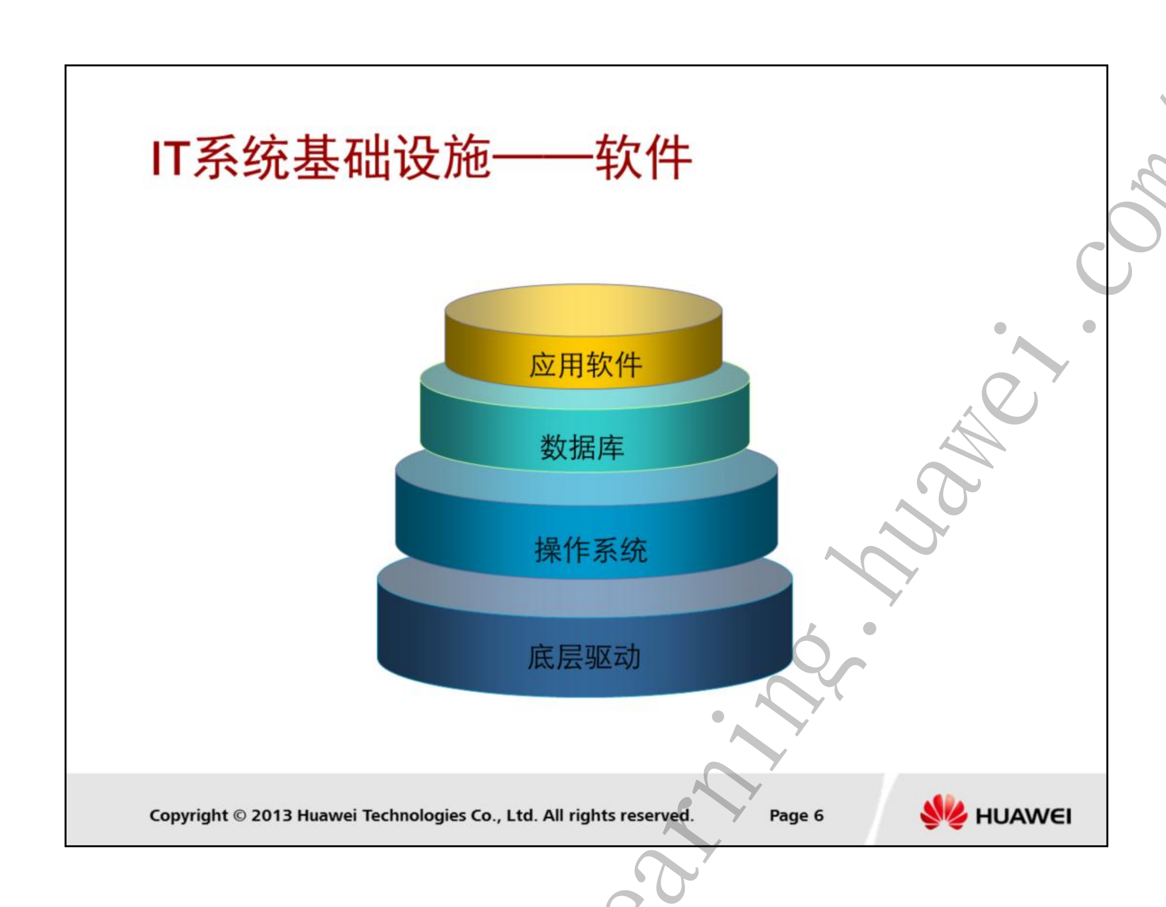
随着大型计算、海量数据存储的发展，对计算能力、数据存储资源方面都有更高的要求，独立的计算机系统已经很难满足。因此，就需要把多个计算机系统集成起来，构成了一个整体的IT系统。IT系统是在计算机系统的基础上所进行的扩展和延伸，其架构上仍然可以划分为计算资源、网络资源、存储资源和软件资源四个层面。

计算资源：在IT系统中，往往会把多台服务器组成集群，通过集群方式实现计算资源的负载均衡，提升整体计算能力，同时，提高系统的冗余，保证系统的可靠性。

存储资源：在IT系统中，存储资源从内置存储向外扩展成为外部存储，通过构建专用的外部存储系统，数据存储得到很大扩展，为大数据量存储提供了必备条件，同时，保证了数据安全可靠。

网络资源：从独立的计算机系统发展成为IT系统，必然需要强大网络资源提供数据通路，计算机系统内部总线已经不能满足IT系统网络资源的要求，因此基于TCP/IP网络和基于FC协议的FC网络架构得到长足的发展，已经成为IT系统网络资源的主流。

软件资源：在IT系统架构中，软件资源不仅仅是独立计算系统中的单一操作系统，而是发展成为集群软件系统、分布式文件系统等，通过这些方式实现集群业务管理和分布式应用。



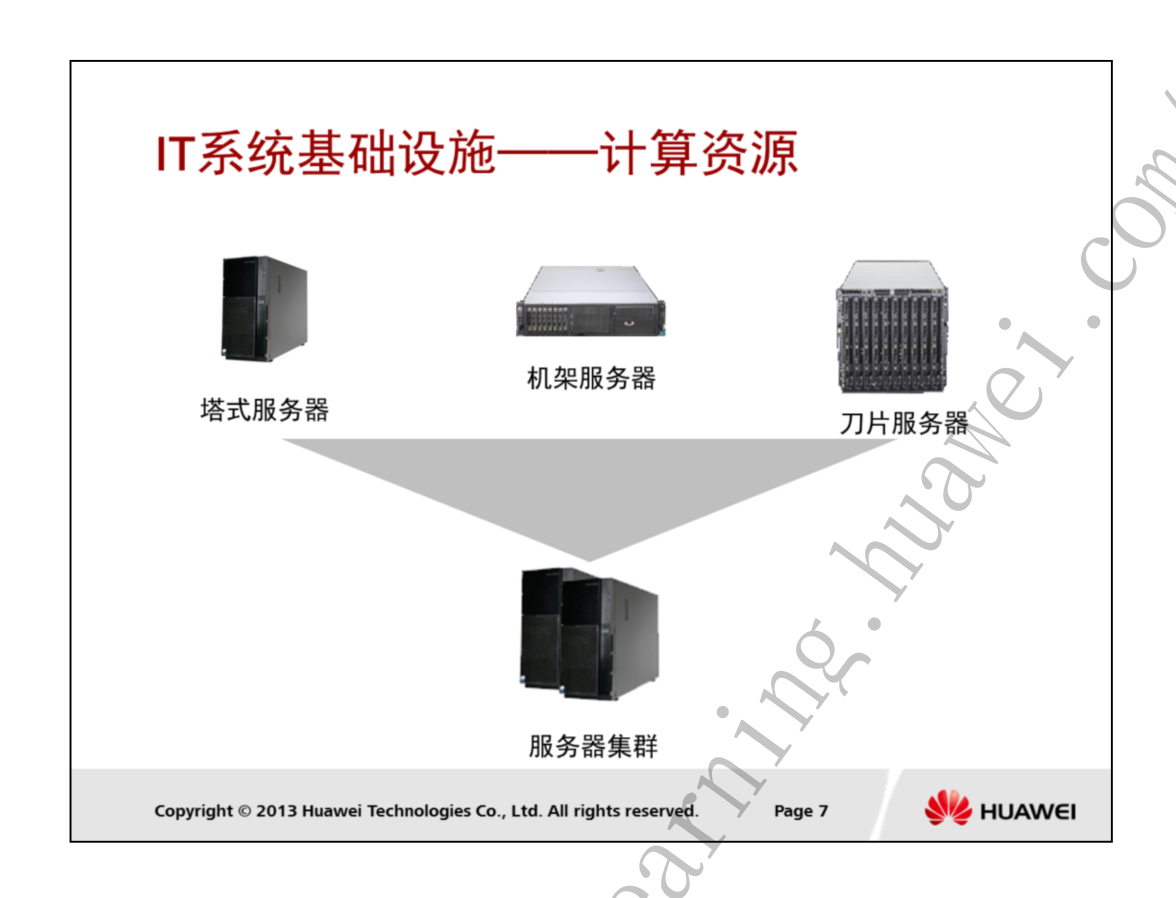
在IT基础设施中，软件也是必不可少的一部分。软件可以从下到上分为：硬件底层驱动、操作系统、数据库、应用软件。

硬件底层驱动和应用软件之间需要实现相应的信息交互。一方面，应用程序通过对驱动程序发送指令，实现硬件控制的动作指令，另一方面，驱动程序将从硬件上获得的数据传送给应用程序，实现应用程序与驱动程序间的交互。也就是说，硬件底层驱动实现了访问底层硬件的人机交互。例如：主板芯片驱动、网卡驱动等。

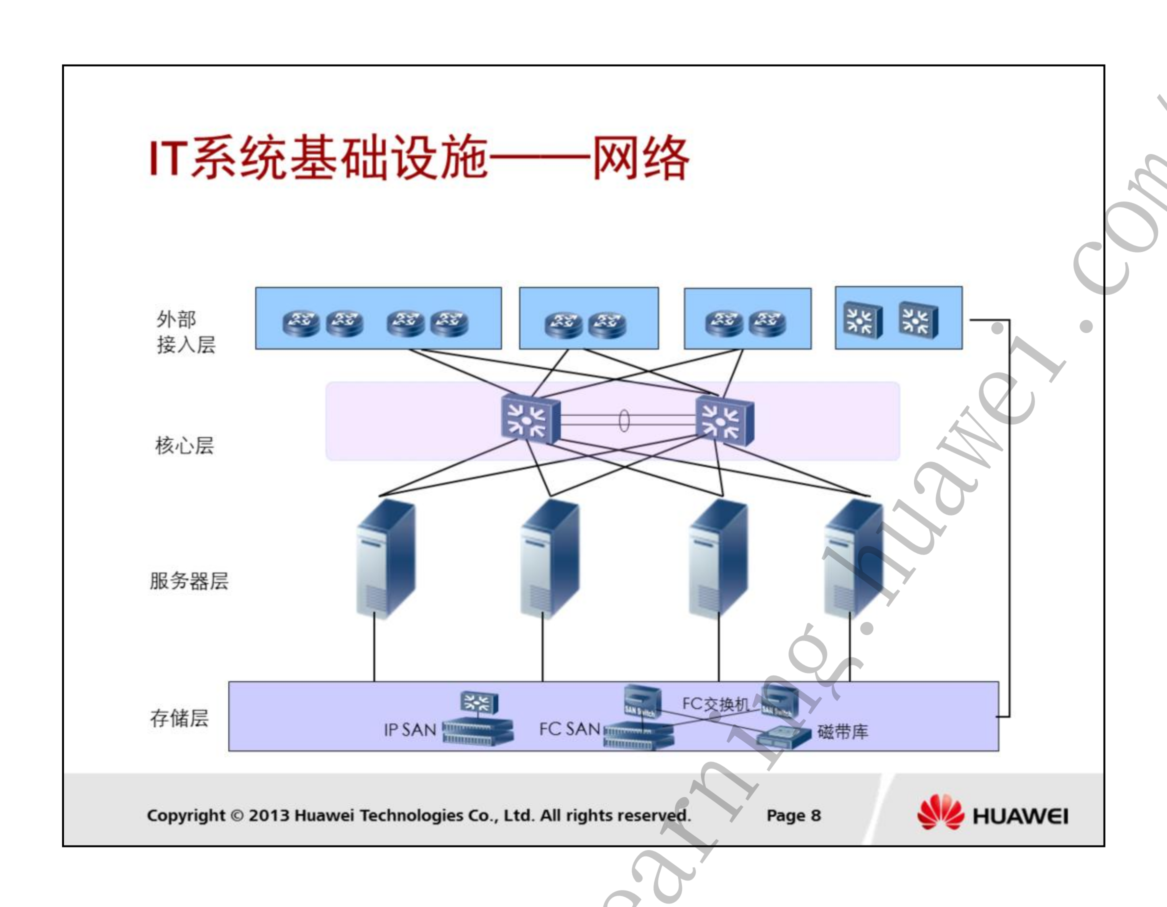
操作系统是管理计算机硬件与软件资源的计算机程序，提供一个让用户与系统交互的操作界面。操作系统需要处理如管理与配置内存、决定系统资源供需的优先次序、控制输入与输出I/O设备、操作网络与管理文件系统等基本事务。例如：微软Windows操作系统、Linux操作系统、Unix操作系统、AIX系统等。

数据库（Database）是按照数据结构来组织、存储和管理数据的仓库。随着信息技术和市场的发展，数据库有发展出很多种类型，从最简单的数据表格到能够进行海量数据存储的大型数据库系统都在各个方面得到了广泛的应用。例如：Oracle、DB2等。

应用软件是为满足用户不同领域、不同业务应用需求而提供的上层软件。 它可以拓宽计算机系统的应用领域，放大硬件的功能。例如：E-mail应用、财务系统等。



IT系统中的计算资源，通常由各式各样的服务器构成。当单个的服务器计算能力不能够满足应用需求的时候，采用服务器集群的形式提供计算能力。

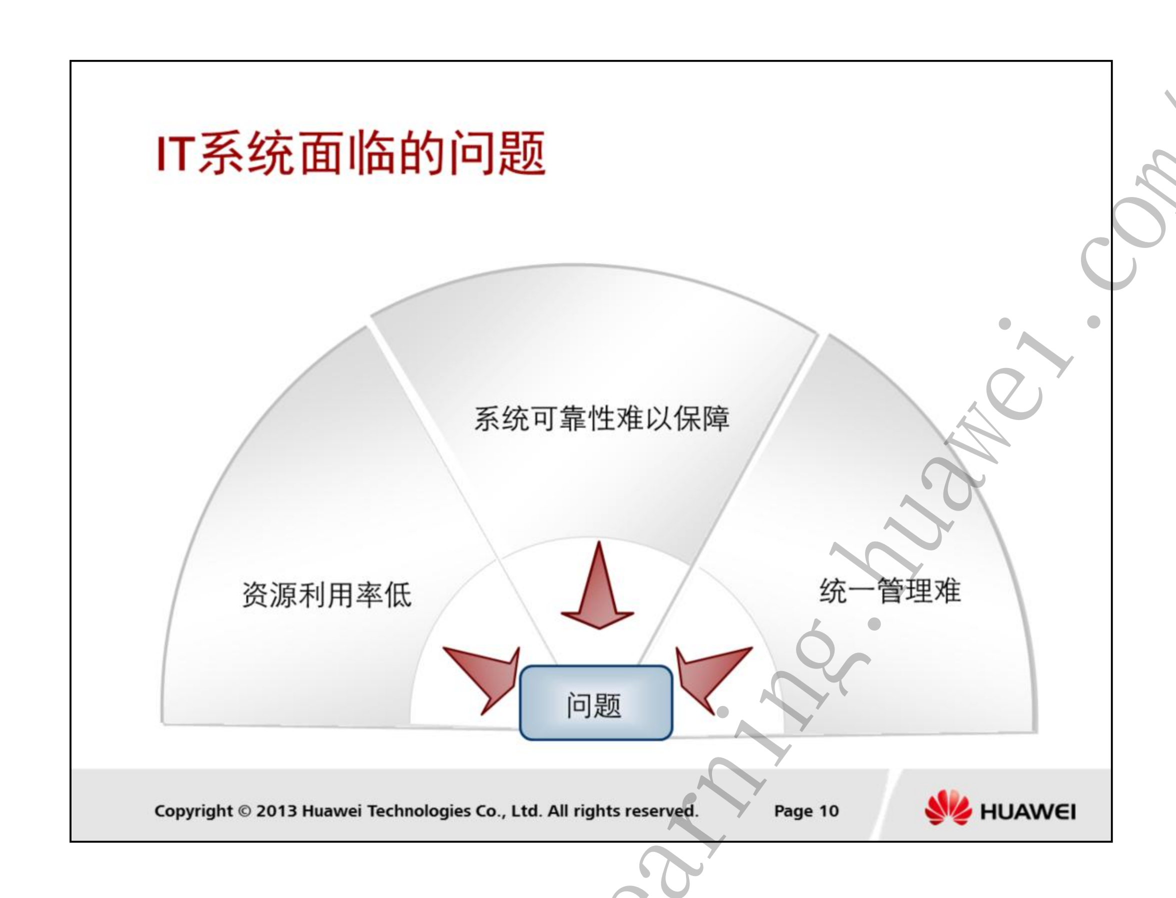


在IT系统基础设施网络架构中主要划分成4个层次，从下到上依次为：存储层、服务器层、核心层、外部接入层。

存储层通过TCP/IP或FC网络连接到服务器层，为服务器提供数据存储空间资源，服务器层接入核心层，内部和外部用户通过接入层、汇聚层连接核心层，在核心层实现快速数据交换。为实现远程容灾，企业需要建立异地灾备中心，通过专用网络实现存储层与灾备中心的互联。



存储资源根据存储的位置可分为内部存储和外部存储，在存储NA后面的课程中，重点介绍外部存储，如SAN存储和NAS设备。

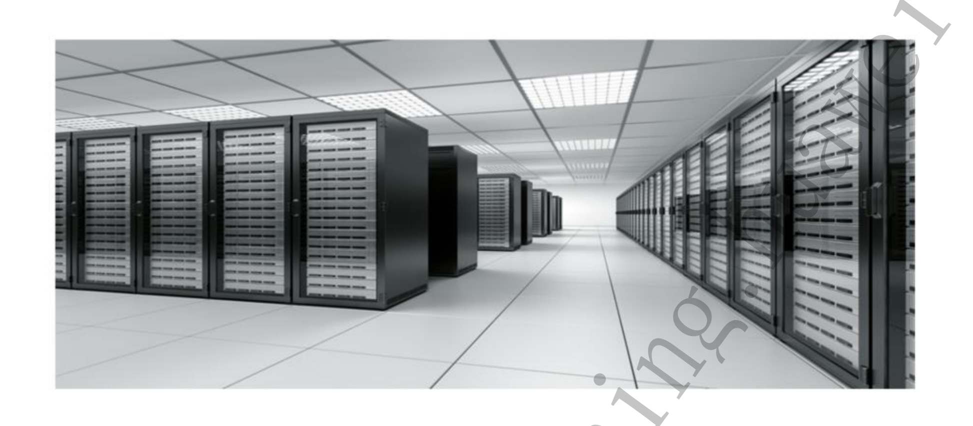


- 资源利用率低：服务器资源，网络资源，存储资源难以得到有效的利用；
- 系统可靠性难以保障：设备的可靠性，传输链路的可靠性难以保障；
- 统一管理难：每一个设备是单独的一套告警系统，要了解整个系统的情况，需要到每一台设备上去查看，每一套设备有单独的密码，管理人员要分别去记，设置复杂了难于记忆，设置简单了存在着一定的安全隐患。

  基于上述原因，出现了数据中心，使以上的问题得到了很好的解决。

数据中心的定义

- 数据中心(DataCenter)通常是指在一个物理空间内实现信息的集中处理、存储、传输、交换、管理。



 Page 11 

计算机设备、服务器设备、网络设备、存储设备等通常认为是数据中心的关键设备。关键设备运行所需要的环境因素，如供电系统、制冷系统、机柜系统、消防系统、监控系统等通常被认为是关键物理基础设施。

数据中心的结构

应用系统
基础办公应用
管理应用
业务应用
基础软件
应用中间件
数据库
操作系统
管理调度
资源管理与调度
基础设施层
物理设备
服务器
存储
网络
机房
空调
UPS
安防

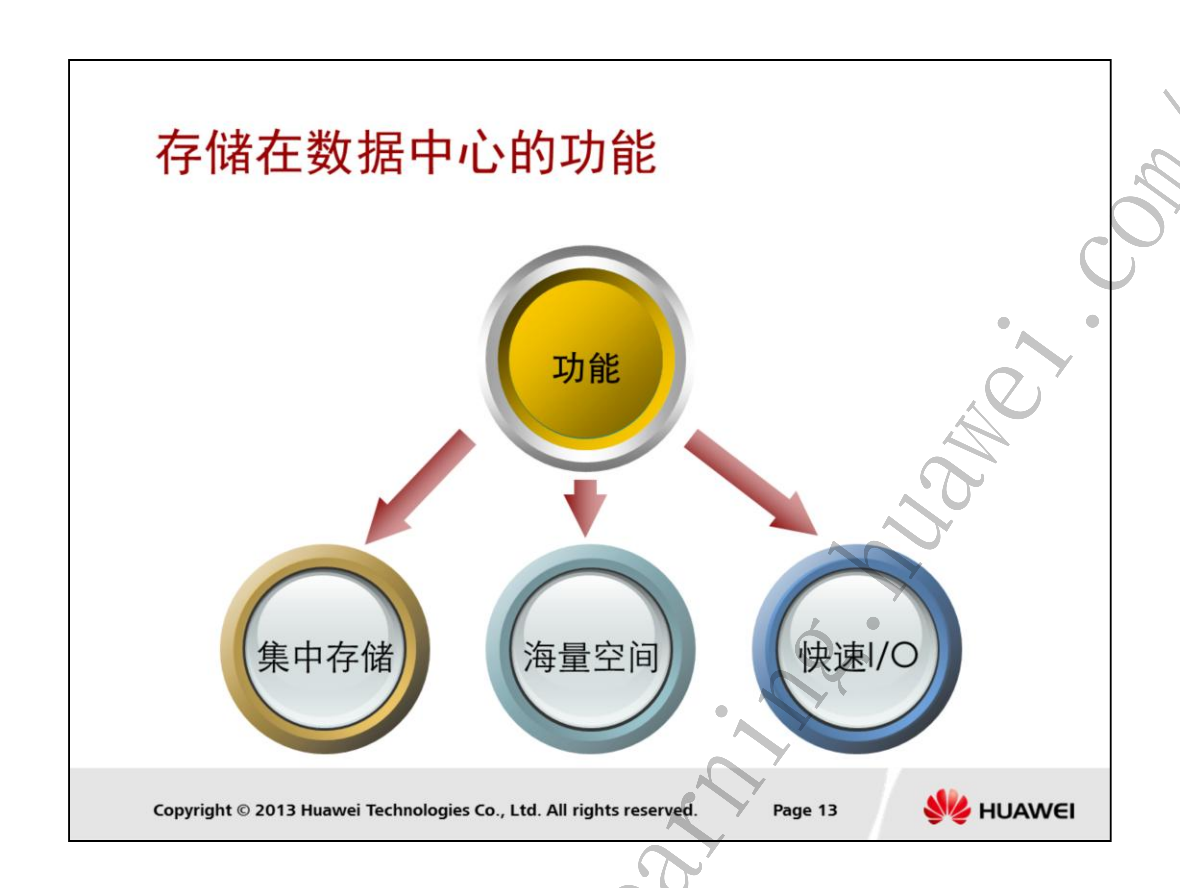
Page 12
HUAWEI

广义的数据中心是企业（机构）的业务系统与数据资源进行集中、集成、共享、分析的场地、工具、流程的有机组合。其核心内容包括业务系统、数据ETL、ODS数据库、数据仓库、数据集市、商务智能等，也包括物理的运行环境（中心机房）和运行维护管理服务。具体来说它包含以下四个方面的具体含义：

1. 数据中心提供所有的应用系统（包括集中的业务应用系统、数据交换平台、应用集成平台）的运营环境；
2. 数据中心是容纳用以支持应用系统运行的基础设施（包括机房、服务器、网络、存储设备）的物理场所；
3. 数据中心包括数据中心本身的ODS、数据仓库及建立在其上的决策分析应用；
4. 数据中心有一套成熟的运行、维护体系支持其日常运行，保证应用系统高效、准确、不间断地运行。

根据数据中心的定义和发展趋势，可以将数据中心划分为4个层次：

1. 基础设施层：用统一的技术将机房、通信、计算、存储等IT基础资源融合形成数据中心的基础设施，为业务系统提供基本的资源服务，提高资源利用率以及IT系统的可靠性。
2. 基础软件层：信息资源是企业生产过程中所涉及的一切文件、资料、图表和数据等信息的总称。本层存储了企业 （机构）生产和经营活动所产生、获取、处理、存储、传输和使用的一切信息资源。
3. 管理调度：实现存储资源化、计算资源化、网络资源化，并能够动态调整资源匹配数据的读写存储，解决统一管理难的问题；
4. 应用层：主要包括针对结构化和非结构化数据的各种应用。包括种业务系统、辅助决策系统和各种多媒体应用（监控、流媒体、统一通信 、呼叫中心、视频 会义 、VOIP）。



- 集中存储

  分散的存储不便于集中管理，存储资源也不利于共享，造成资源的浪费。

- 海量空间

  随着Internet的发展，海量的非结构化的数据的诞生，对数据中心的存储也提出了新的要求，在一定程度上，海量存储空间的大小直接决定着数据中心的发展规模。

- 快速I/O

  由于数据中心的用户众多，I/O的快速响应能力直接决定着用户的应用感知，因此，快速的I/O也为数据中心的高效运行提供了保障。



目录



主机的内部应用环境

- 主机内部I/O流程各个环节共同构成了数据存储的内部应用环境。

上层业务应用
I/O请求
操作系统
调用接口
文件系统
地址映射
设备驱动
SCSI指令
存储设备

 Page 15 HUAWEI

主机服务器大部分I/O开始于需要访问数据的应用，应用通常不考虑存储后端的操作细节，而是直接调用由操作系统提供的系统调用接口，然后由操作系统支持的文件系统为数据提供数据的逻辑地址和在磁盘上存储的物理地址的映射，再通过设备驱动层，主要是SCSI协议的操作，将数据存储到存储设备（比如主机服务器内部硬盘）上。因此主机服务器内部数据存储I/O流程中的各个环节对数据存储的可靠性、性能和安全性都起到非常重要作用，从独立的主机服务器来看，其内部I/O流程各个环节即共同构成了存储的内部应用环境。

传统内置存储遇到的问题

- 硬盘成为整个系统的性能瓶颈
  - 有限的硬盘槽位，难满足大容量需求
  - 单个硬盘存放数据，数据可靠性难以保证
- 存储空间利用率低
  - 本地存储，数据分散，难以共享
- 可扩展性不够
  - 总线结构，而非网络结构
  - 可连接的设备受到限制增加容量时，需停机

传统内置存储

接口 接口 接口

处理器

内存

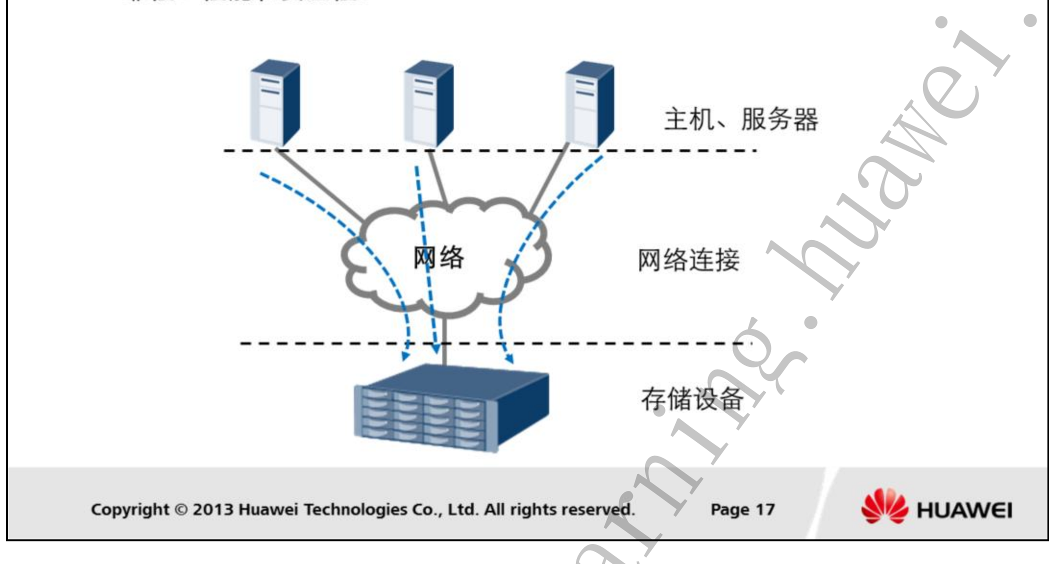


硬盘 硬盘

 Page 16 HUAWEI

在传统的计算机存储系统中，存储工作通常是由计算机内置的硬盘来完成，而采用这样的设计方式，硬盘本身的缺陷很容易成为整个系统的性能瓶颈，并且，由于机箱内有限的空间，限制了硬盘数量的扩展，并且同时也对机箱内的散热、供电等提出了严峻的挑战。再加上不同的计算机各自为战，使用各自内置的硬盘，导致从总体看来存储空间的利用率较低，并且分散保存的数据也不利于数据的共享和备份工作。在传统的C/S架构中，无论使用的是何种协议，存储设备都直接与服务器相连接。在这样的结构下，对存储设备上所保存的所有数据的任何读写操作，都必须由服务器来进行，这样的处理方式给服务器带来了沉重的负担。

外部存储系统的出现，彻底将服务器从繁琐的I/O操作中解放出来，使服务器更加专门化，使之仅仅承担应用数据的操作任务，以更充分的释放自身潜能。



网络存储系统按照不同的功能，可划分为三层：

第一层：主机层，运行相关应用，发起存储IO操作。在主机侧需要存储连接的设备主要有FC HBA卡，iSCSI HBA卡，以太网卡，需在主机侧安装的存储连接软件包括：initiator启动器,open-iscsi,多路径软件等。

第二层：网络层，将主机与存储层互联，提供网络通路，可采用FC网络或者以太网络的方式；以太网络可利用原有以太网络连接，利用现有资源组网。FC网络需要建立专门的网络，速度快，但是FC器件都较昂贵，成本高。

第三层：存储层，存储系统的核心层，对数据进行存储、管理。

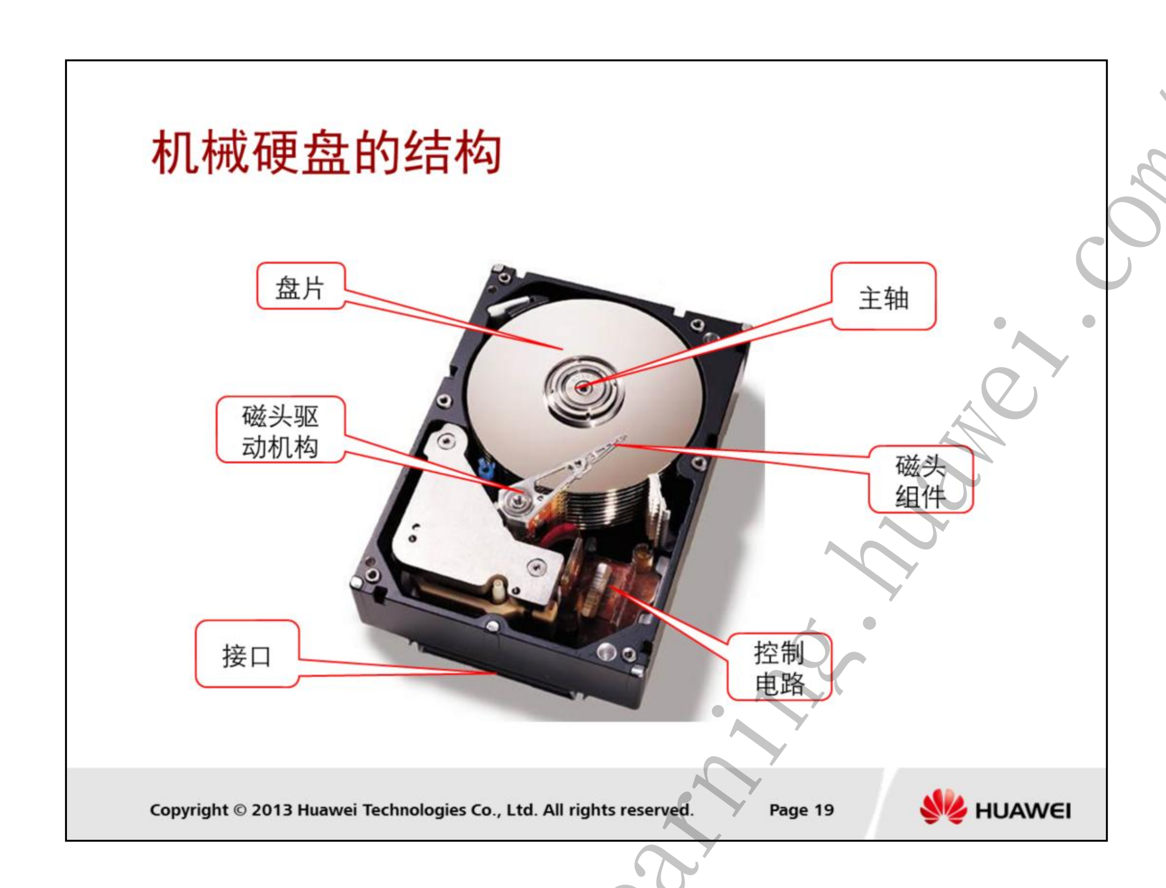
上面提到的每一层都将对数据存储的可靠性、性能、安全性产生重要影响，因此在进行存储系统建设时，需要从以上各层使用的硬件设备、软件协议、组网架构等方面考虑，以保证业务应用对数据存储可靠性、性能、安全性方面的需求。

下面我们将从存储介质和主机应用维度对存储应用环境影响因素进行介绍，涉及到存储网络及存储设备部分的内容将在后面的章节中重点介绍。



目录





机械硬盘包含机械装置和电子装置，可以分为如下部分：

- 磁头组件

  用于数据的读取和写入。

- 磁头驱动机构

  用于驱动磁头臂将磁头送达指定的位置。

- 盘片组

  数据的载体。

- 主轴驱动装置

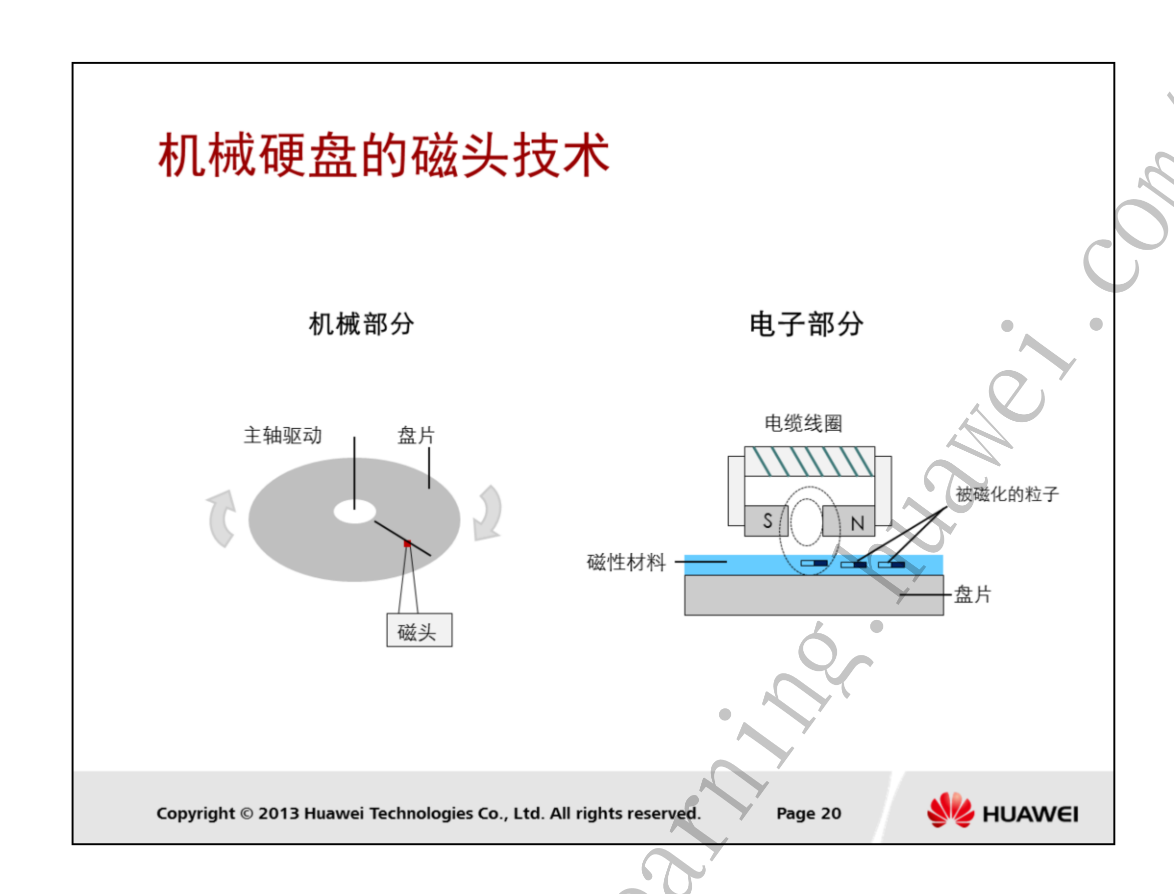
  驱动盘片维持高速运转。

- 控制电路

  系统控制、调速、驱动等。

- 接口

  用于硬盘与主板连接，常见的接口类型有ATA,SATA,SAS。



- 机械部分

  1. 系统在密封机构里

  2. 盘片固定并由主轴驱动进行高速旋转

  3. 磁头沿盘片径向移动

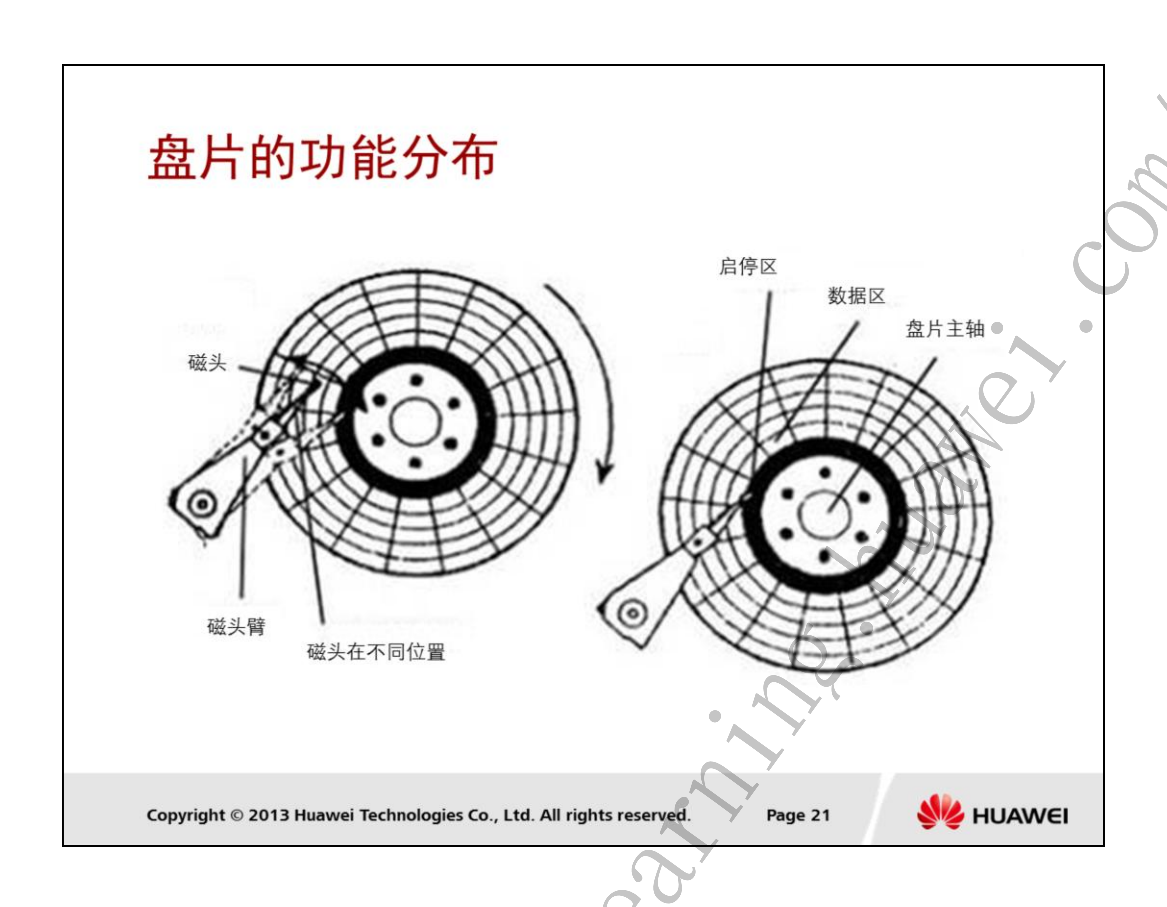
  4. 磁头在盘片上方飞行

- 电子部分

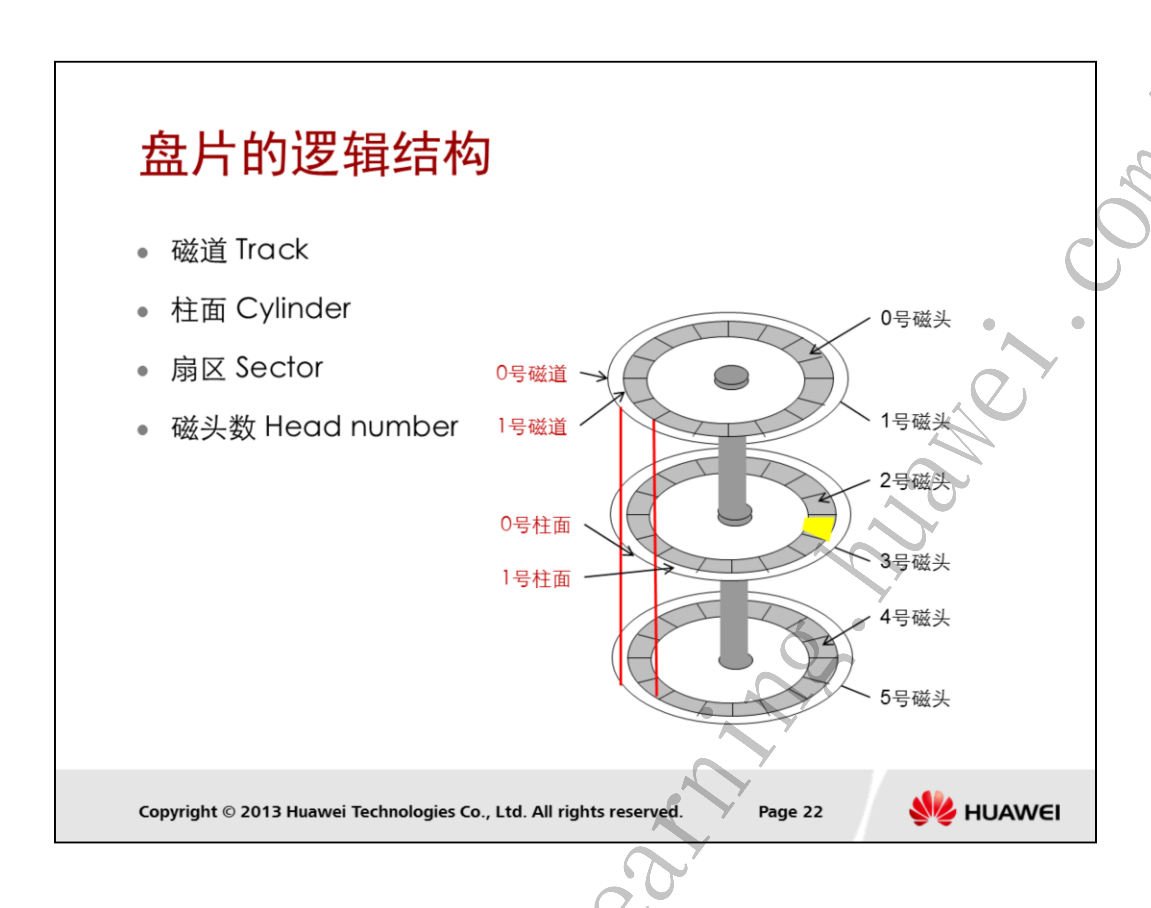
  1. 盘片上溅镀金属性粒子，呈不规则排列

  2. 通过控制线圈上的电流，磁头形成磁场

  3. 对盘面上的金属粒子进行磁化(整齐排列)



磁头靠近主轴接触的表面，即线速度最小的地方，是一个特殊的区域，它不存放任何数据，称为启停区或着陆区（Landing Zone），启停区外就是数据区。在最外圈，离主轴最远的地方是“0”磁道，硬盘数据的存放就是从最外圈开始的。那么，磁头是如何找到“0”磁道的位置的呢？在硬盘中还有一个叫“0”磁道检测器的构件，它是用来完成硬盘的初始定位。



- 磁道 Track

磁盘上人为规定的若干个同心圆的轨道称为磁道；磁道从外向内由0开始编号。300~1024，甚至更多。

- 柱面 Cylinder

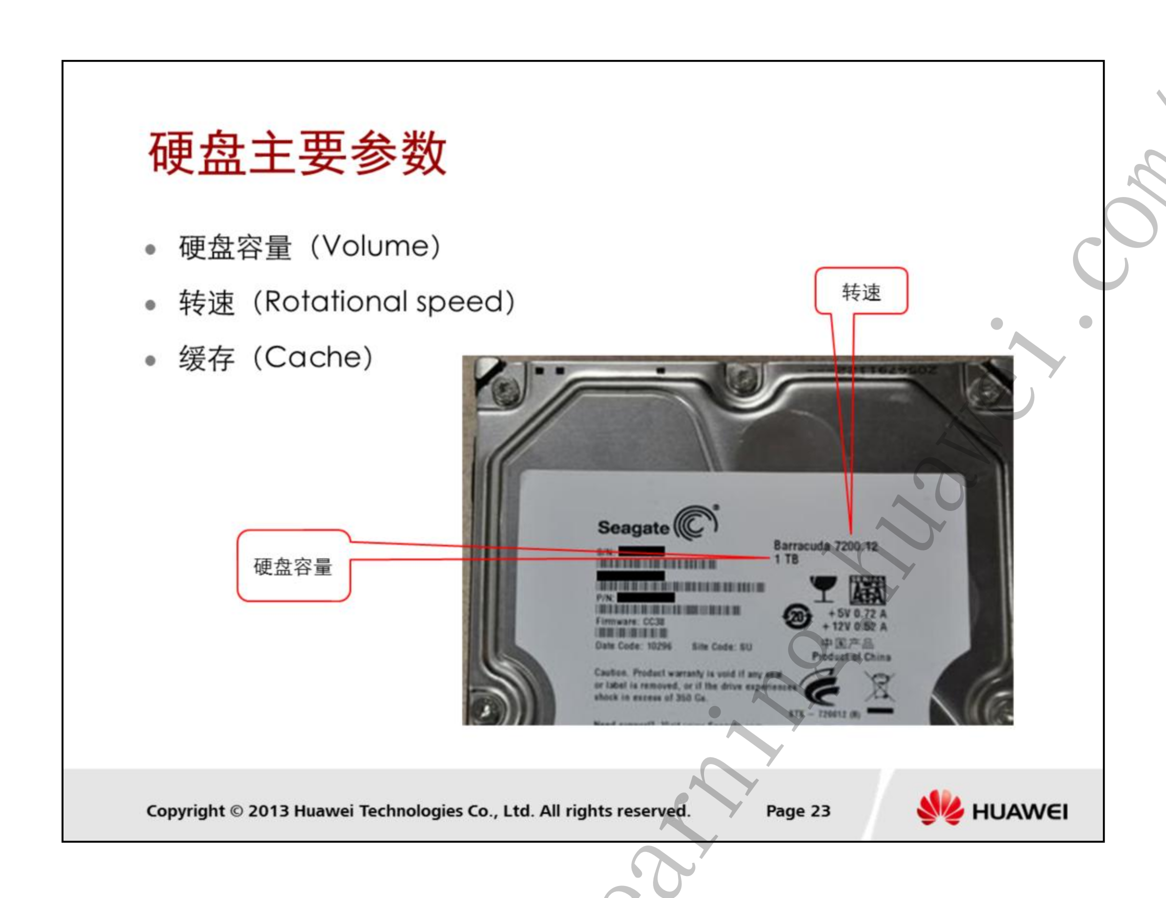
所有盘面上的同一磁道构成一个圆柱，称为柱面；柱面从外向内由0开始编号，和磁道数目一致；

- 扇区 Sector

磁盘上每个磁道被分为若干个弧段，由1开始编号；每个弧段可以存储512或4K字节的信息，称为扇区。

- 磁头数 Head number

磁盘上每个盘面都有对应的读写磁头；磁头数与盘面数一致。



- 硬盘容量（Volume）

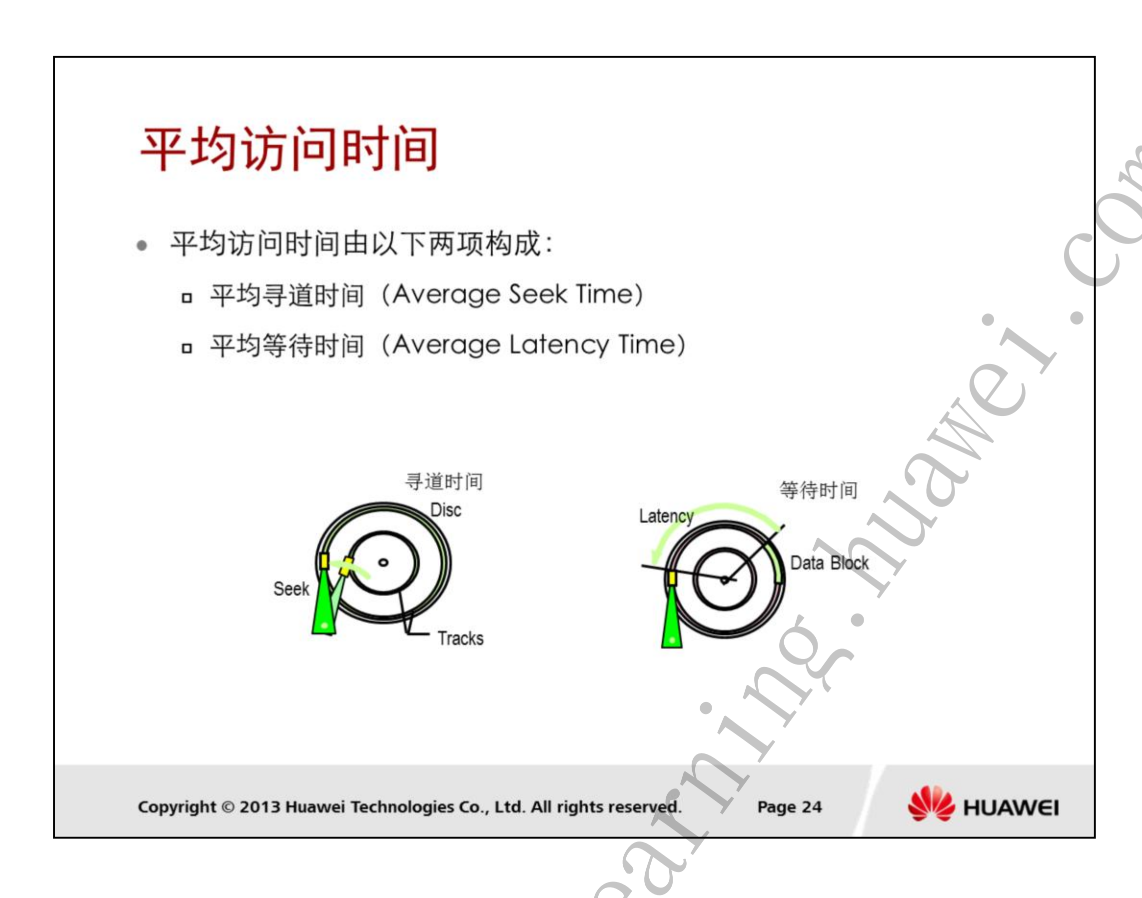
容量的单位为兆字节（MB）或千兆字节（GB）。影响硬盘容量的因素有单碟容量和碟片数量。

- 转速（Rotational speed）

硬盘的转速是指硬盘盘片每分钟转过的圈数，单位为RPM（Rotation Per Minute）。一般硬盘的转速都达到5400RPM/7200RPM。有些SCSI接口的硬盘使用了液态轴承技术，转速可达10000－15000RPM。

- 缓存（Cache）

由于CPU与硬盘之间存在巨大的速度差异，为解决硬盘在读写数据时CPU的等待问题，在硬盘上设置适当的高速缓存，以解决二者之间速度不匹配的问题。硬盘缓存与CPU上的高速缓存作用一样，是为了提高硬盘的读写速度。



- 平均寻道时间（Average Seek Time）

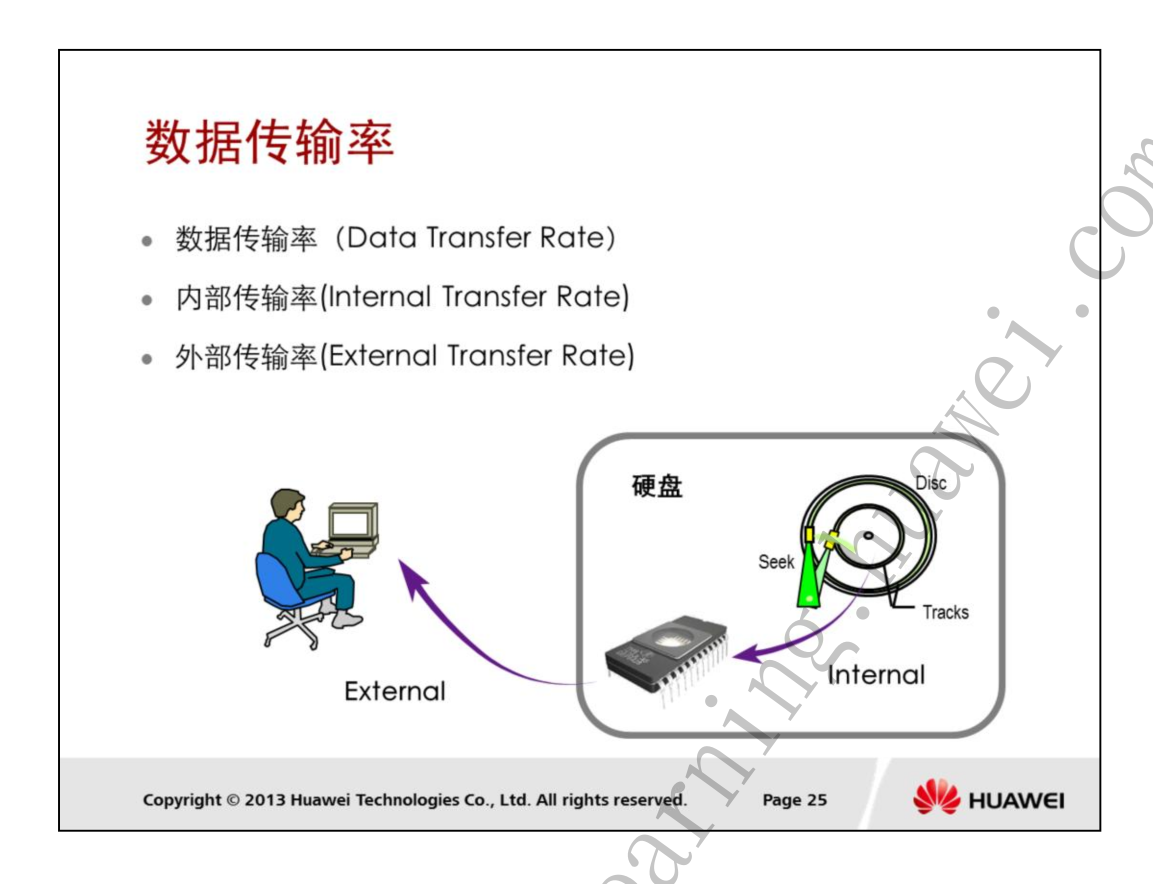
硬盘的平均寻道时间是指硬盘的磁头从初始位置移动到盘面指定磁道所需的时间，是影响硬盘内部数据传输率的重要参数。这个时间越小越好。目前IDE硬盘的平均寻道时间通常在 8ms到11ms之间。

- 硬盘的等待时间（Average Latency Time）

硬盘的等待时间又叫潜伏期，是指磁头已处于要访问的磁道，等待所要访问的扇区旋转至磁头下方的时间。平均等待时间通常为盘片旋转一周所需时间的一半，因此硬盘转速越快，等待时间就越短，一般应在4ms以下。

- 平均访问时间（Average Access Time）

平均访问时间 = 平均寻道时间 + 平均等待时间。



- 数据传输率（Data Transfer Rate）

硬盘的数据传输率是指硬盘读写数据的速度，单位为兆字节每秒（MB/s）。硬盘数据传输率包括内部传输率和外部传输率两个指标。

- 内部传输率(Internal Transfer Rate)

内部传输率也称为持续传输率（Sustained Transfer Rate），是缓存之间的数据传输速度，它反映了硬盘缓冲区没有使用时的性能，这可以说是影响硬盘整体速度的瓶颈。内部传输率主要依赖于硬盘的磁头到硬盘的高速旋转速度。值得注意的是Mb/S或Mbps为单位，这是兆位/秒

- 外部传输率(External Transfer Rate)

外部传输率也称为突发数据传输率（Burst Data Transfer Rate）或接口传输率。它指的是系统总线与硬盘缓冲区之间的数据传输率，与硬盘接口类型和硬盘缓存的大小有关。

IOPS和Throughput

- IOPS
  - IOPS (Input/Output Per Second)即每秒的输入输出量(或读写次数)，是衡量磁盘性能的主要指标之一。
- Throughput
  - Throughput吞吐量,指单位时间内可以成功传输的数据数量。对于大量顺序读写的应用，如电视台的视频编辑，视频点播VOD(Video On Demand)，则更关注吞吐量指标。

 Page 26 

- IOPS计算方法

磁盘完成一个I/O请求所花费的时间，它由寻道时间、旋转延迟和数据传输时间三部分构成。

寻道时间Tseek是指将读写磁头移动至正确的磁道上所需要的时间。寻道时间越短，I/O操作越快，目前磁盘的平均寻道时间一般在3 – 15ms。

旋转延迟Trotation是指盘片旋转将请求数据所在扇区移至读写磁头下方所需要的时间。旋转延迟取决于磁盘转速，通常使用磁盘旋转一周所需时间的1/2表示。比如，7200 rpm的磁盘平均旋转延迟大约为60*1000/7200/2 = 4.17ms，而转速为15000 rpm的磁盘其平均旋转延迟约为2ms。

数据传输时间Ttransfer是指完成传输所请求的数据所需要的时间，它取决于数据传输率，其值等于数据大小除以数据传输率。目前IDE/ATA能达到133MB/s，SATA II可达到300MB/s的接口数据传输率，数据传输时间通常远小于前两部分时间。因此，理论上可以计算出磁盘的最大IOPS，即IOPS = 1000 ms/ (Tseek + Troatation)，忽略数据传输时间。假设磁盘平均物理寻道时间为3ms, 磁盘转速为7200,10K,15K rpm，则磁盘IOPS理论最大值分别为，

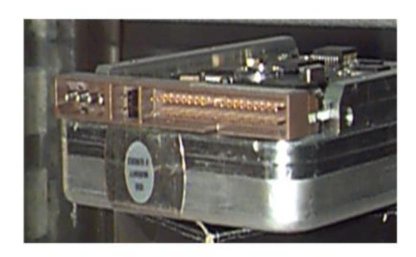
IOPS = 1000 / (3 + 60000/7200/2) = 140

IOPS = 1000 / (3 + 60000/10000/2) = 167

IOPS = 1000 / (3 + 60000/15000/2) = 200

硬盘常用接口——ATA接口

- ATA(Advanced Technology Attachment) 高级技术附加装置
  - ATA硬盘也经常称为IDE Integrated Drive Electronics 硬盘
  - ATA接口为并行ATA技术





 Page 27

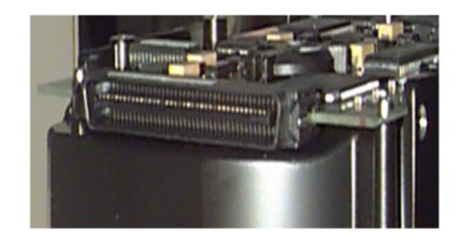


ATA接口发展到今，细分可以分成ATA-1（IDE）、ATA-2（EIDE Enhanced IDE/Fast ATA）、ATA-3（FastATA-2）、Ultra ATA、Ultra ATA/33、Ultra ATA/66、Ultra ATA/100及Serial ATA。ATA发展到ATA100已经到了尽头，转向SATA。

- ATA接口具有以下优缺点：
  - 优点： 价格低廉、兼容性非常好；
  - 缺点： 速度慢、只能内置使用、对接口电缆的长度有很严格的限制

硬盘常用接口—— SCSI接口

- SCSI (Small Computer System Interface) 小型计算机系统接口。





 Page 28



SCSI硬盘并发处理性能优异，常应用于企业级存储领域。SCSI硬盘分50、68、80针，由SCSI-1 不断发展至当前主流Ultra 320（320MB/s）。

- 优点：
  - 适应面广，在一块SCSI控制卡上就可以同时挂接15个设备；
  - 高性能（具有多任务、宽带宽及CPU占用率低等特点）；
  - 具有外置和内置两种，支持热插拔。
- 缺点：
  - 价格昂贵，安装复杂。

硬盘常用接口—— SATA接口

- SATA：Serial ATA (Serial ATA) 串行ATA
  - SATA采用串行方式进行数据传输，接口速率比IDE接口高，最低为150MBps，并且第二代（SATAⅡ ）300MBps接口硬盘已经形成商用，规划内的最高速率可达600MBps
  - SATA硬盘采用点对点连接方式，支持热插拔，即插即用



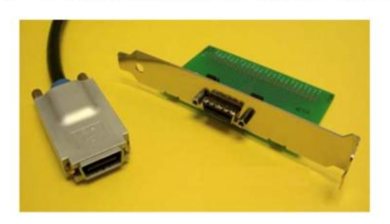
 Page 29



SATA接通常为7+15针，采用单通道，传输速率要比ATA更快。SATA具有比ATA更好的抗干扰能力。

硬盘SAS接口

- SAS(Serial Attached SCSI) 串行连接SCSI
  - SAS是一种点对点、全双工、双端口的接口
  - SAS专为满足高性能企业需求而设计，实现与SATA的互操作，为企业用户带来前所未有的灵活性和低成本
  - 速率每路600M
  - SAS具有高性能、高可靠性、强大的扩展性能





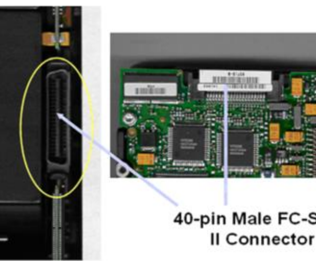
Page 30

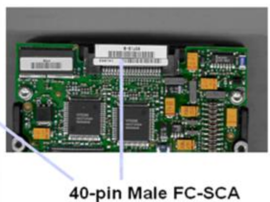


SAS可以向下兼容SATA，同样采用串行技术，在传输速率、抗干扰性方面强于SCSI，SAS接口硬盘价格相对更高。

硬盘常用接口—— FC接口

- FC硬盘采用FC-AL(Fiber Channel Arbitrated Loop) 光纤通道仲裁环
  - FC-AL是一种双端口的串行存储接口
  - FC-AL支持全双工工作方式
  - FC-AL利用类似SATA/SAS所用的4芯连接，提供一种单环拓扑结构，一个控制器能访问126个硬盘





40-pin Male FC-SCA
II Connector



 Page 31 

FC硬盘定位于高端存储应用，可靠性和性能高，FC硬盘一般都同时提供两个FC接口，可同时使用或互为备份。

优点：

- 具有很好的升级性，可以用非常长的光纤电缆，可超过10公里；
- 具有非常高的带宽；
- 具有很强的通用性。

缺点：

- 价格非常昂贵；
- 组件复杂。

硬盘常用接口—— NL SAS

- NL SAS 采用SAS接口、SATA 盘体，也叫近线SAS。

类别	时效性	容量	性能	访问速度	成本
在线	即时服务	小	高	快	高
近线	非即时的	较大	低	较快	低
离线		大	低	慢	低

 Page 32



NL SAS 是指采用了SAS接口和SATA盘体的综合体。即具有SAS接口、接近SAS性能的SATA盘。NL是Near Line的缩写，意为近线。所谓的近线存储，主要定位于客户在线存储和离线存储之间的应用。就是指将那些并不是经常用到，或者说数据的访问量并不大的数据存放在性能较低的存储设备上。但同时对这些的设备要求是寻址迅速、传输率高。（例如客户一些长期保存的不常用的文件的归档）。因此，近线存储对性能要求相对来说并不高，但又要求相对较好的访问性能。同时多数情况下由于不常用的数据要占总数据量的比较大的比重，这也就要求近线存储设备需要容量相对较大。

在线存储(OnStore)是工作级的存储，在线存储的最大特征是存储设备和所存储的数据时刻保持“在线”状态，可以随时读取和修改，以满足前端应用服务器或数据库对数据访问的速度要求。

离线存储(OffStore)主要使用光盘或磁带存储。大多数情况下主要用于对在线存储的数据进行备份，以防范可能发生的数据灾难，因此又称备份级的存储。

SAS、NL-SAS与SATA的应用比较

	SAS	NL-NAS	SATA
优势	高可靠性 高性能 原生支持SCSI 支持双端访问 高级容错技术	原生支持SCSI 支持双端访问 高级容错技术 大容量 低功耗	大容量 低功耗
推荐场景	业务量大 访问频率较高 以小数据块居多 数据较为离散的高/中端用户	更适合大数据块 业务压力不大的用户使用	适合大数据块 业务压力不大的用户使用

 Page 33

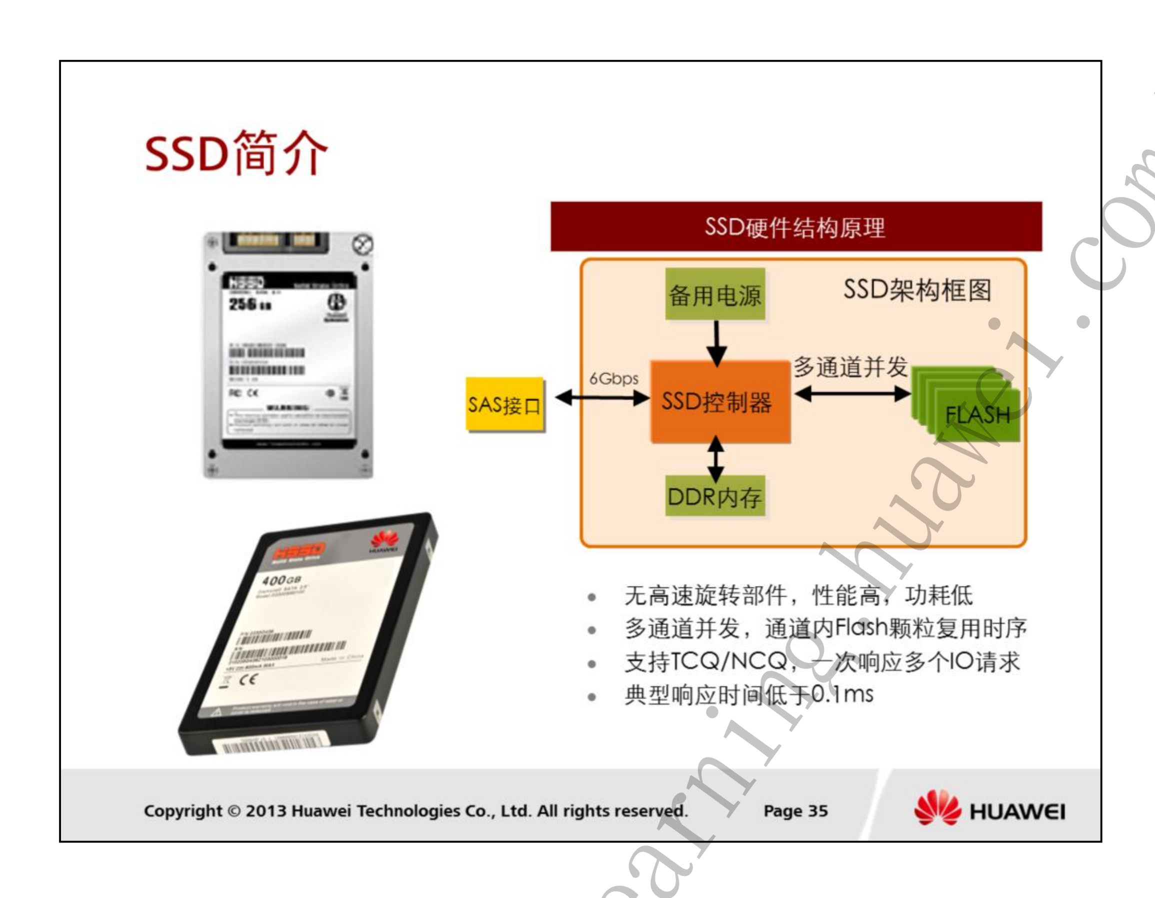


支持双端访问：双端访问的目的主要不是为了提高硬盘的访问带宽，而是为了硬盘的高可用性和容灾的需要而采取的一种技术，在存储系统中，硬盘的双端口分别接到存储的A控，B控，当某一个控制器上的端口出现故障的时候，可通过另一个端口访问。



目录



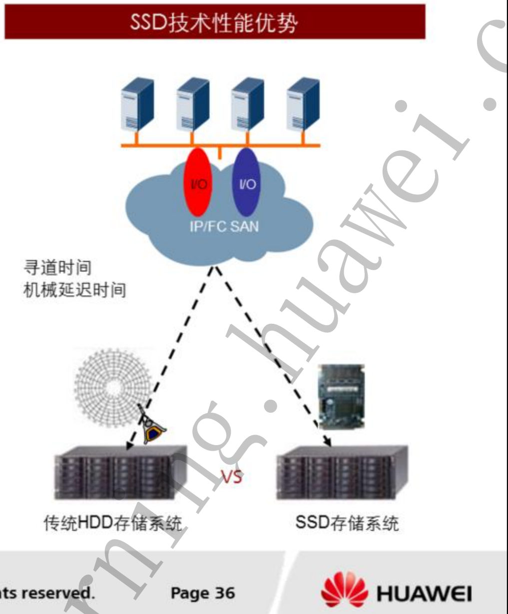


NCQ（Native Command Queuing）与TCQ（Tagged Command Queuing）都是设计通过把计算机发向硬盘的指令做重新排序，从而提高硬盘性能的技术。NCQ技术在 300MB/s的Serial ATA II规格中引入，针对的是主流的硬盘产品，而TCQ技术是在SCSI2规格中引入（ATA-4标准中也有采用），针对的是服务器以及企业级硬盘产品。

要使用NCQ、TCQ技术，芯片组硬盘接口和硬盘产品本身都必须支持才行，也就是说，如果你购买的一款新硬盘并不支持NCQ，即使你的主板是最新的支持NCQ的，也不能够打开这个功能从而提高性能。

SSD性能优势

- 响应时间短
  - 机械硬盘的机械特性导致大部分时间浪费在寻道和机械延迟上，数据传输效率受到严重制约。
- 读写效率高
  - 机械硬盘在进行随机读写操作时，磁头不停地移动，导致读写效率低下，而SSD通过内部控制器计算出数据的存放位置，直接进行存取操作，故效率高。



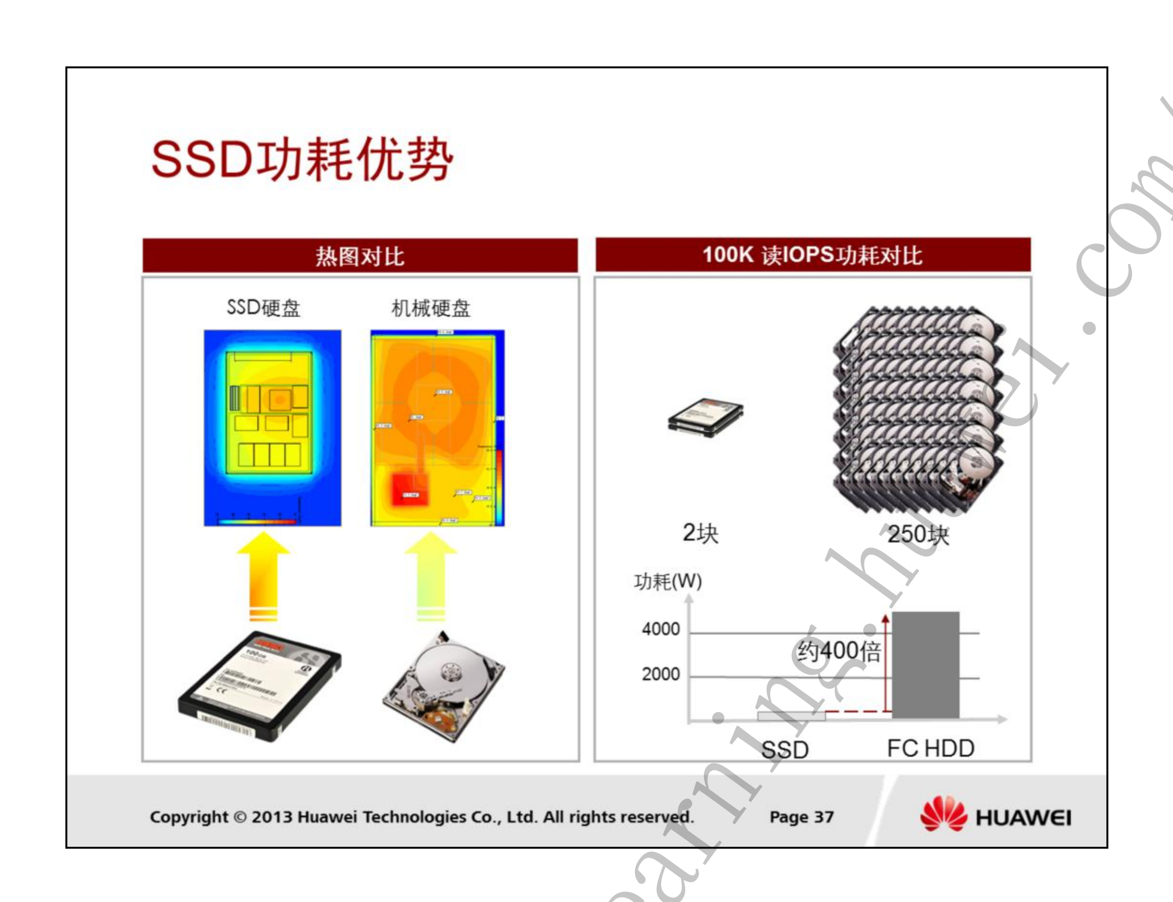
 Page 36 HUAWEI

- 响应时间短

传统硬盘的机械特性导致大部分时间浪费在寻道和机械延迟上，数据传输效率受到严重制约。而SSD硬盘内部没有机械运动部件，省去了寻道时间和机械延迟，可更快捷的响应读写请求

- 读写效率高

机械硬盘在进行随机读写操作时，磁头不停地移动，导致读写效率低下，而SSD通过内部控制器计算出数据的存放位置，并进行读写操作， 省去了机械操作时间，大大提高了读写效率。4K随机读写情况下：FC硬盘的性能为400/400 IOPS，SSD硬盘的性能为26000/5600 IOPS

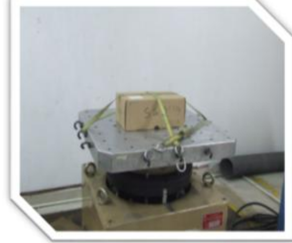


较少的硬盘功耗优势不是很明显，但是当使用大量的磁盘数量的时候，功耗所产生的成本差别就比较大，也成为企业在选择方案时候的一个考虑因素。

SSD的环境适应优势

- SSD不含高速旋转的机械结构部件，可经得住严苛的环境考验，以华为SSD硬盘为例：
  - HSSD可承受振动加速度16.4G，机械硬盘一般为0.5G以下
  - HSSD抗冲击1500G，机械硬盘一般为70G左右
- HSSD使用专用设备做过如下测试：
  - 静压试验、跌落试验、随机振动试验、冲击试验、碰撞试验

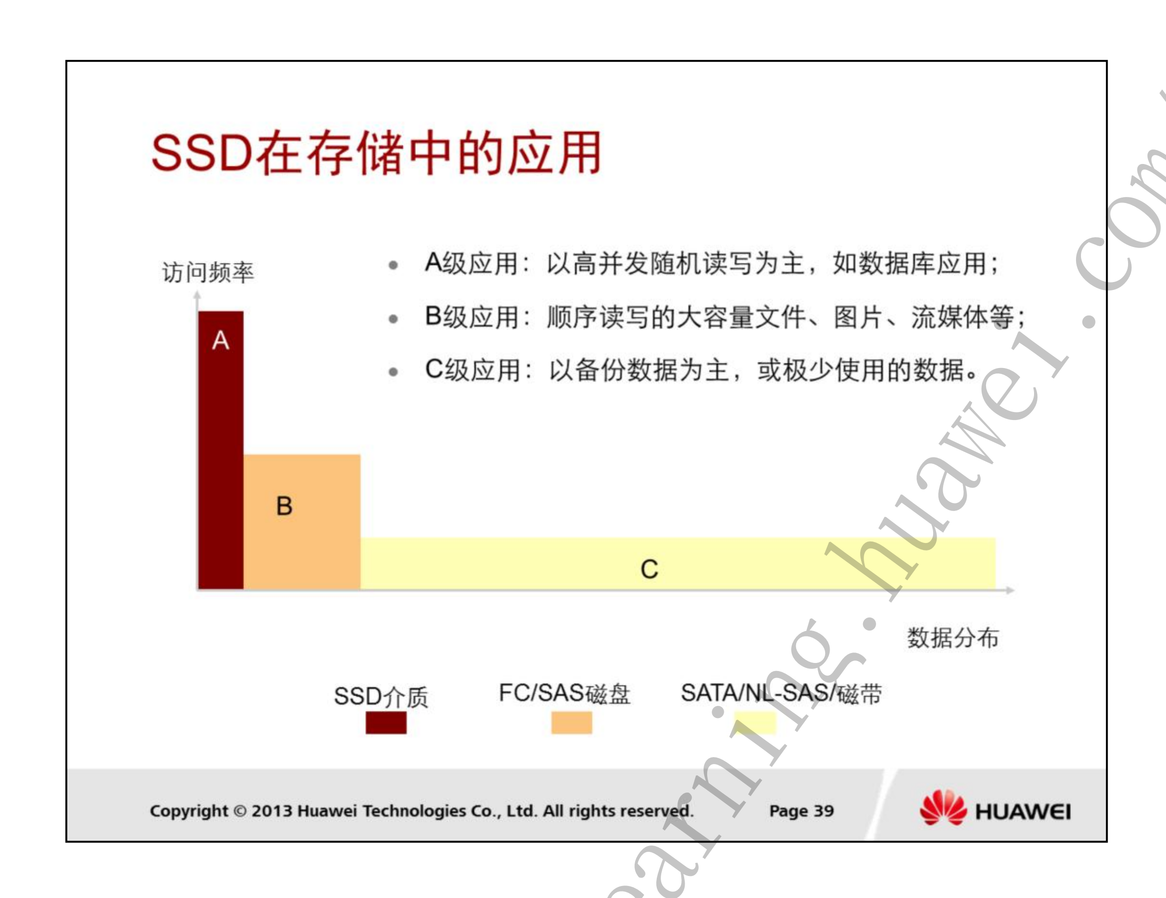





 Page 38



SSD可用在一些环境较恶劣的场合，如高温高湿、强震等恶劣环境下使用，比如很多工业级应用要求SSD固态硬盘做到-20到70摄氏度和-40到85摄氏度的宽温要求。

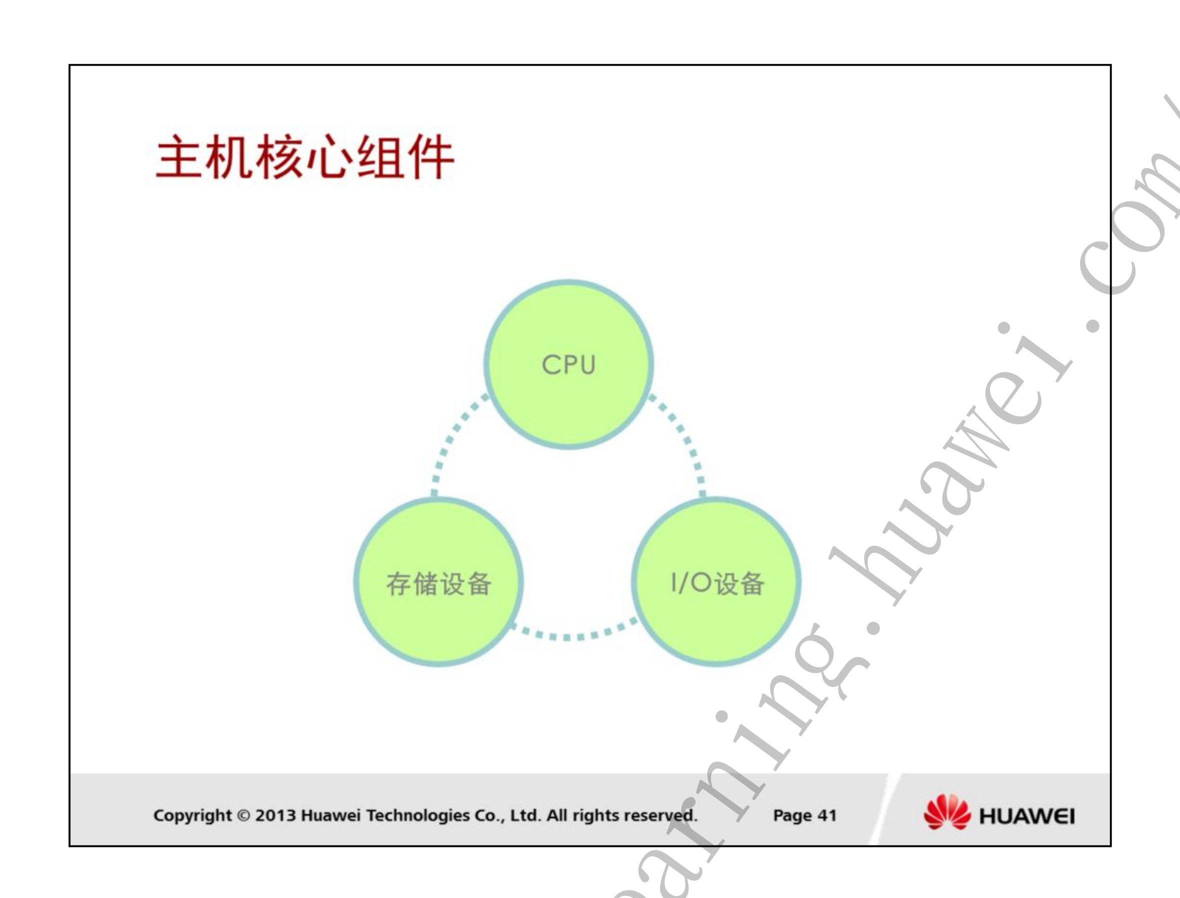


- 二八原则
  - 用户需要频繁改动或读写的数据，一般占存储总量的20%，称为热数据，对应于A级应用。
- 分级存储
  - 将热数据存放在SSD上，B级和C级应用的数据存放在高速HDD和一般HDD上，提升性能并减少投资。

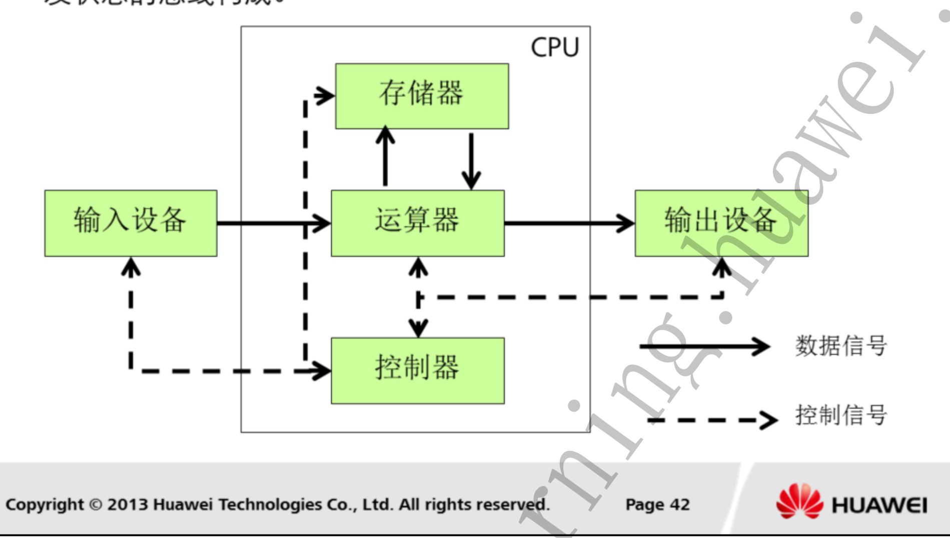


目录





主机的核心组件由CPU，存储设备，I/O设备三个部分组成，这三个部分通过总线互联通讯。



CPU的运作原理可分为四个阶段：提取（Fetch）、解码（Decode）、执行（Execute）和写回（Writeback）。 CPU从存储器或高速缓冲存储器中取出指令，放入指令寄存器，并对指令译码，并执行指令。

CPU的指令集

- CICS指令集
  - Intel X86架构的CPU
- RISC指令集
  - PowerPC处理器、SPARC处理器
- EPIC指令集
  - Intel的安腾Itanium CPU

 Page 43



- CISC指令集

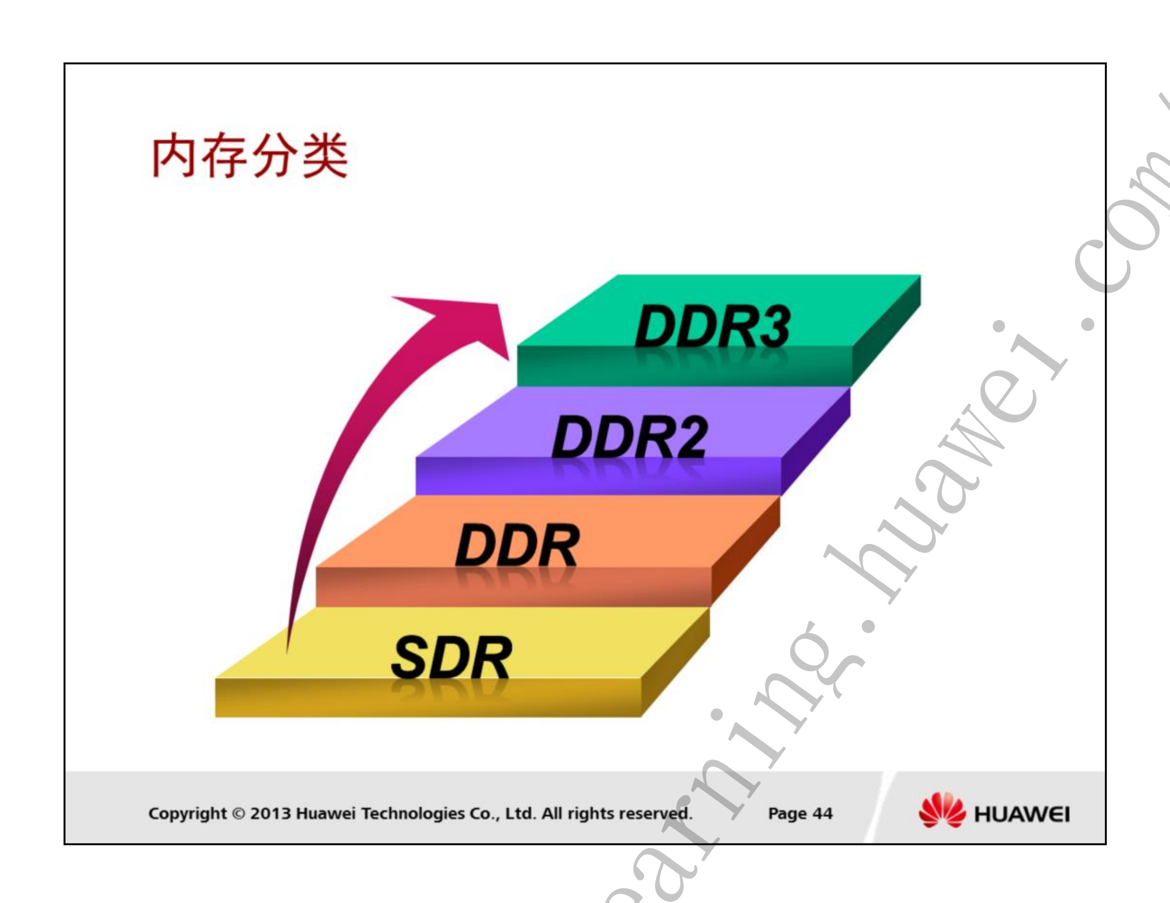
CISC（Complex Instruction Set Computer）指令集，也称为复杂指令集。英特尔生产的x86系列CPU及其兼容CPU，如AMD、VIA都是属于CISC的范畴。由于Intel X86系列及其兼容CPU都使用X86指令集，所以就形成了今天庞大的X86系列及兼容CPU阵容。

- RISC指令集

RISC（Reduced Instruction Set Computing ）即“精简指令集”。它是在CISC指令系统基础上发展起来的。RISC指令系统更加适合高档服务器的操作系统UNIX，Linux也属于类似UNIX的操作系统。RISC型CPU与Intel和AMD的CPU在软件和硬件上都不兼容。在中高档服务器中采用RISC指令的CPU主要有以下几类：PowerPC处理器、SPARC处理器、PA-RISC处理器、MIPS处理器、Alpha处理器。

- EPIC 指令集

EPIC（Explicitly Parallel Instruction Computers，精确并行指令计算机）可以说是Intel的处理器迈向RISC体系的重要步骤。Intel采用EPIC技术的服务器CPU是安腾Itanium，也是IA-64系列中的第一款。微软也已开发了64位的操作系统，在软件上加以支持。



SDR（synchronous）同步动态随机存取存储器，这种内存的特性是在一个内存时钟周期中，在一个方波上升沿时进行一次操作(读或写),一般有两个缺口；

DDR（DOUBLE DATA RATE）： SDRAM的更新换代产品，他允许在时钟脉冲的上升沿和下降沿传输数据，这样不需要提高时钟的频率就能加倍提高SDRAM的速度。DDR内存不向后兼容SDRAM。

DDR2采用和DDR1内存一样的指令,工作频率比DDR1更高，DDR2的缺口比较靠中间；

DDR3与DDR2的基础架构并没有本质的不同，DDR3能提供更高的外部数据传输率，更先进的地址/命令与控制总线的拓朴架构，在保证性能的同时将能耗进一步降低。

目前普及的是DDR3内存,主要参数有容量、主频、带宽。

存储设备

硬盘：是电脑主要的存储媒介之一，由一个或者多个铝制或者玻璃制的碟片组成

磁带：是一种用于记录声音、图像、数字或其他信号的载有磁层的带状材料，是产量最大和用途最广的一种磁记录材料

光盘：光盘以光信息做为存储物的载体

Flash存储：一种不挥发的内存器件


硬盘


磁带

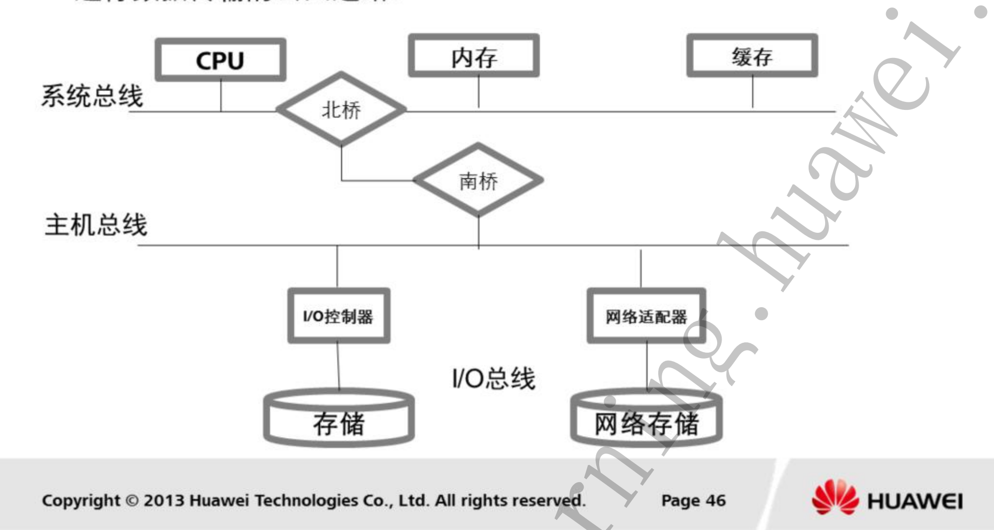

光盘


Flash存储

 Page 45



- 硬盘
  - 寻址访问、数据存储速度快，成本高
  - 适合做快速响应访问的场合
- 磁带
  - 顺序读写、读写速度快、容量大、脱机存放容易、成本低
  - 适合做长期保存、快速读写的场合
- 光盘
  - 寻址访问、保存简单、可靠性高、低成本
  - 适合做长期保留，对写速度要求不高的场合
- Flash存储
  - 能长久保持数据，容量可观，携带方便
  - 适合做长期保存，对读速度要求不高的场合



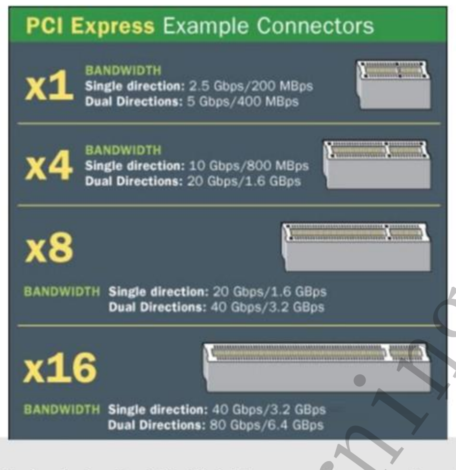
CPU通过系统总线读取内存中的指令，通过指令的计算结果对内存进行读写操作，CPU和内存之间通过高速的北桥芯片互联。

主机总线指一些适配器连接到南桥的总线，比如显卡，网卡，RAID卡，FC HBA卡等

I/O总线指适配器与外部设备相连的总线，比如硬盘，U盘，存储等。

什么是PCI-E总线?

- Intel 在2001年春季的IDF上，正式公布了旨在取代PCI总线的第三代I/O 技术，最后却被正式命名为PCI- Express ，Express 意思是高速、特别快的意思。



 Page 47



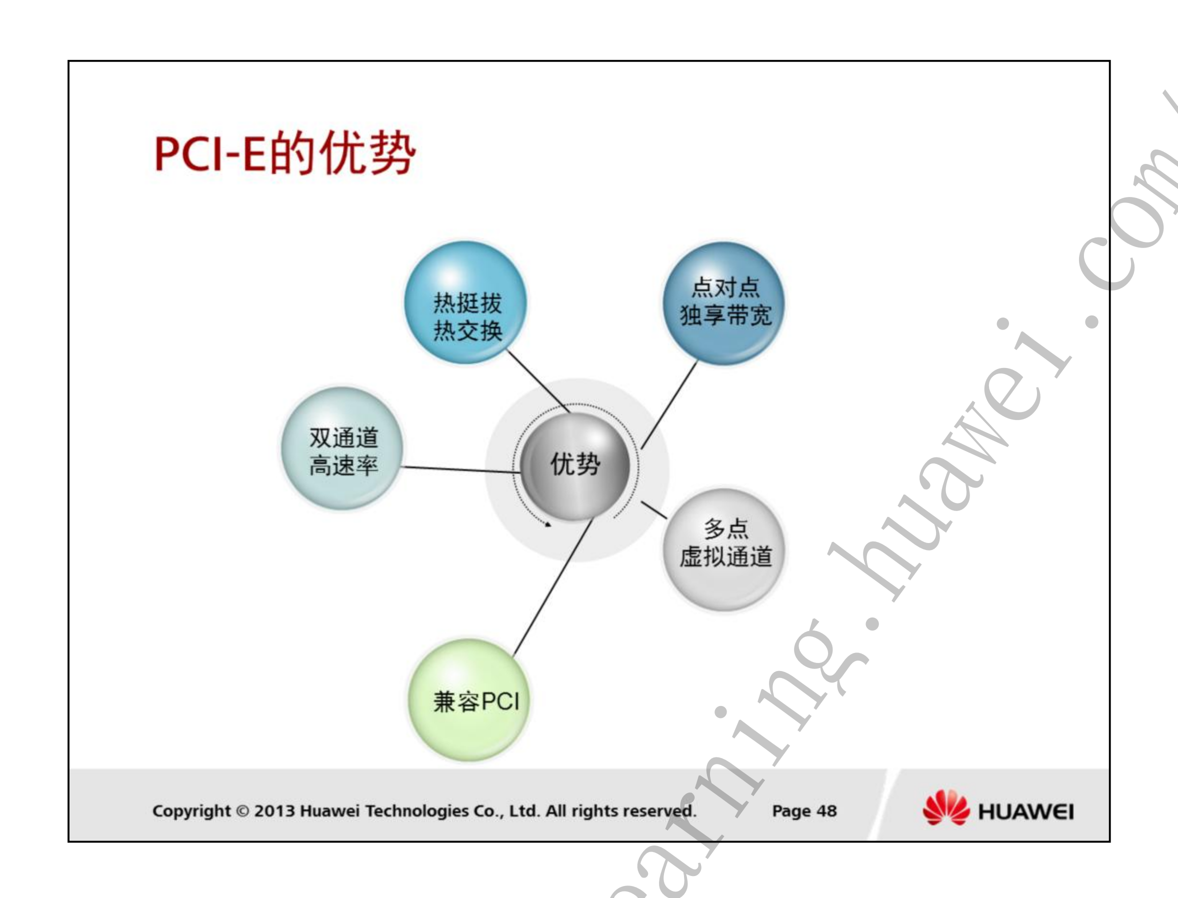
随着计算机技术的发展，PCI 总线已经无法满足电脑性能提升的要求，必须由带宽更大、适应性更广、发展潜力更深的新一代总线取而代之，这就是PCI-Express 总线。

Intel 在2001年春季的IDF上，正式公布了旨在取代PCI总线的第三代I/O 技术，最后却被正式命名为PCI- Express ，Express 意思是高速、特别快的意思。

2002年7月23日，PCI-SIG 正式公布了PCI Express 1.0 规范，并且根据开发蓝图，在2006年的时候正式推出PCI Express2.0规范。

经历着这么三代半（AGP总线只是一种增强型的PCI总线）的发展，PC的外部总线终于发展到我们现在看到的PCI-E 2.0，提供了比以往总线大得多的带宽。

目前的主流应用是X8和X16。



- 点对点独享带宽：

　　PCI是所有设备共享同一条总线资源，PCI-E采用点对点技术，芯片之间用接口连线，设备之间用数据电缆。

- 双通道，高速率：

　　类似于全双工模式，速度大提升，1.0版本的PCI-E在每个信道单方向2.5Gpbs的传输速率作为起步，物理上可达到1-32个可选信道；

- 热插拔热交换：

　　PCI-E总线接口插槽中含有“热拔插检测信号”，所以可以像USB总线那样进行热插拔和热交换；

- 多点虚拟通道：

　　类似于InfiniBand，PCI-E总线技术在每一个物理通道中也支持多点虚拟通道，理论上每一个单物理通道中可以有 8 条虚拟通道进行独立的通信控制，而且每个通信的数据包都可以定义不同的QoS。

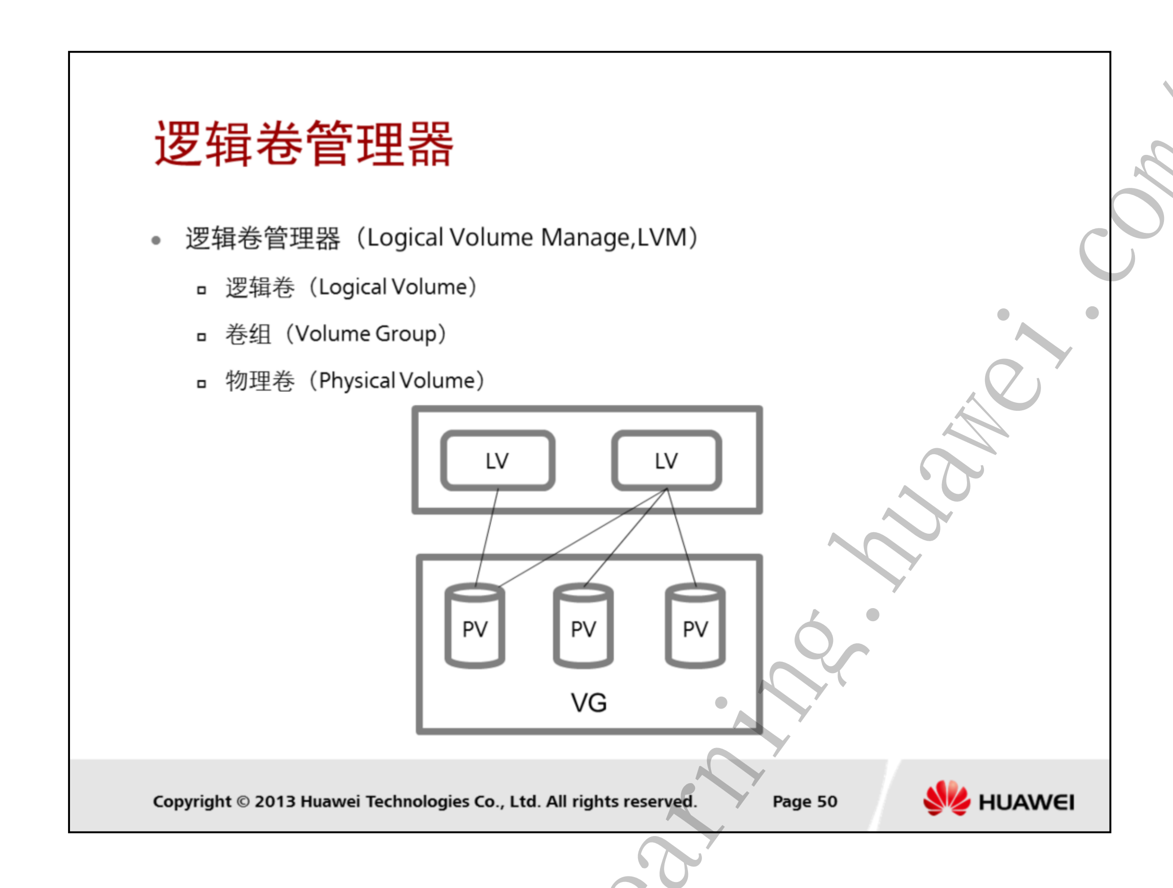
- 兼容PCI:

　　以前的PCI可以在PCI-E的这一模式下运行，为用户提供了一个平滑的升级平台，但要注意的是不兼容目前的AGP接口。



目录





在主机软件系统中，操作系统上层的应用软件用于满足用户不同领域、不同业务应用需求，操作系统实现对计算机硬件与软件资源管理，并提供用户与系统交互的操作界面。不同的操作系统使用其各自的文件系统，文件系统建立在逻辑卷的基础上。

- 逻辑卷

逻辑卷改变了数据的存储方式，使数据的存储更具有灵活性，一个逻辑卷可以跨越多个物理磁盘，数据存储在物理上可以是不连续的。逻辑卷是建立在卷组之上的,卷组中的空间可以建立多个逻辑卷,并且逻辑卷可以随意从卷组的空闲空间中增减,逻辑卷可以属于一个卷组,也可以属于不同的多个卷组。

- 卷组

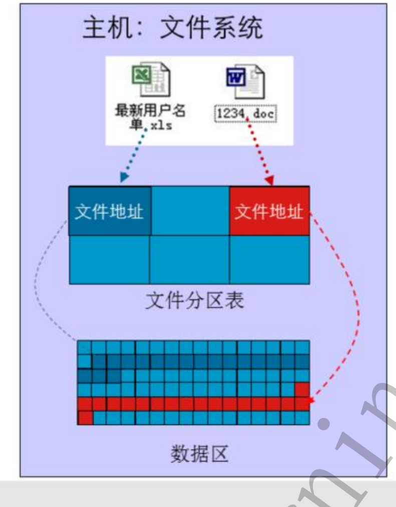
卷组是建立在物理卷之上,一个卷组中可以包含一个物理卷组或者多个物理卷。

- 物理卷

物理卷在逻辑卷管理器中属于最底层的,任何的逻辑卷和卷组都必需依靠物理卷来建立,物理卷可以是一个完整的硬盘,也可以是硬盘中的莫一个分区。

文件系统

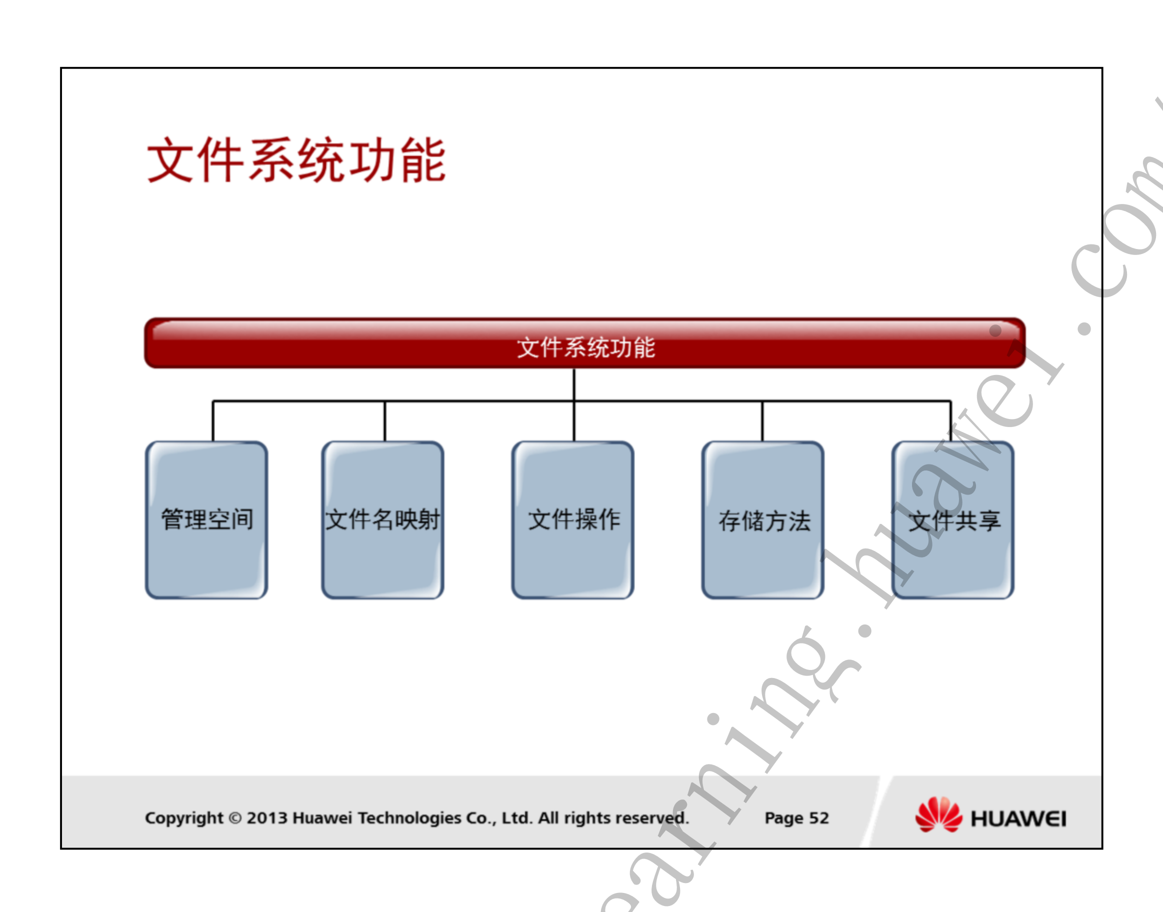
- 文件系统：是指把文件存储于硬盘时所必须的数据结构及硬盘数据的管理方式。



 Page 51 

为了访问硬盘中的数据，就必需在扇区之间建立联系，也就是需要一种逻辑上的数据存储结构。建立这种逻辑结构就是文件系统要做的事情，在硬盘上建立文件系统的过程通常称为“格式化”。

硬盘数据的管理通过文件分区表，记录数据的地址，然后通过地址记录实现对数据的读取。



文件系统的功能包括：管理和调度文件的存储空间，提供文件的逻辑结构、物理结构和存储方法;实现文件从标识到实际地址的映射(即按名存取)，实现文件的控制操作和存取操作(包括文件的建立、撤销、打开、关闭，对文件的读、写、修改、复制、转储等)，实现文件信息的共享并提供可靠的文件保密和保护措施，提供文件的安全措施(文件的转储和恢复能力)。

操作系统的主要功能

- 进程管理（Processing management）
- 内存管理（Memory management）
- 网络通讯（Networking）
- 安全机制（Security）
- 用户界面（User interface）
- 驱动程序（Device drivers）



- 进程管理

  管理进程间的通讯，进程异常终止以及进程死锁等；

- 内存管理

  提供寻找可用内存空间，配置与释放内存空间等；

- 网络通讯

  提供通讯协议的协商；

- 安全机制

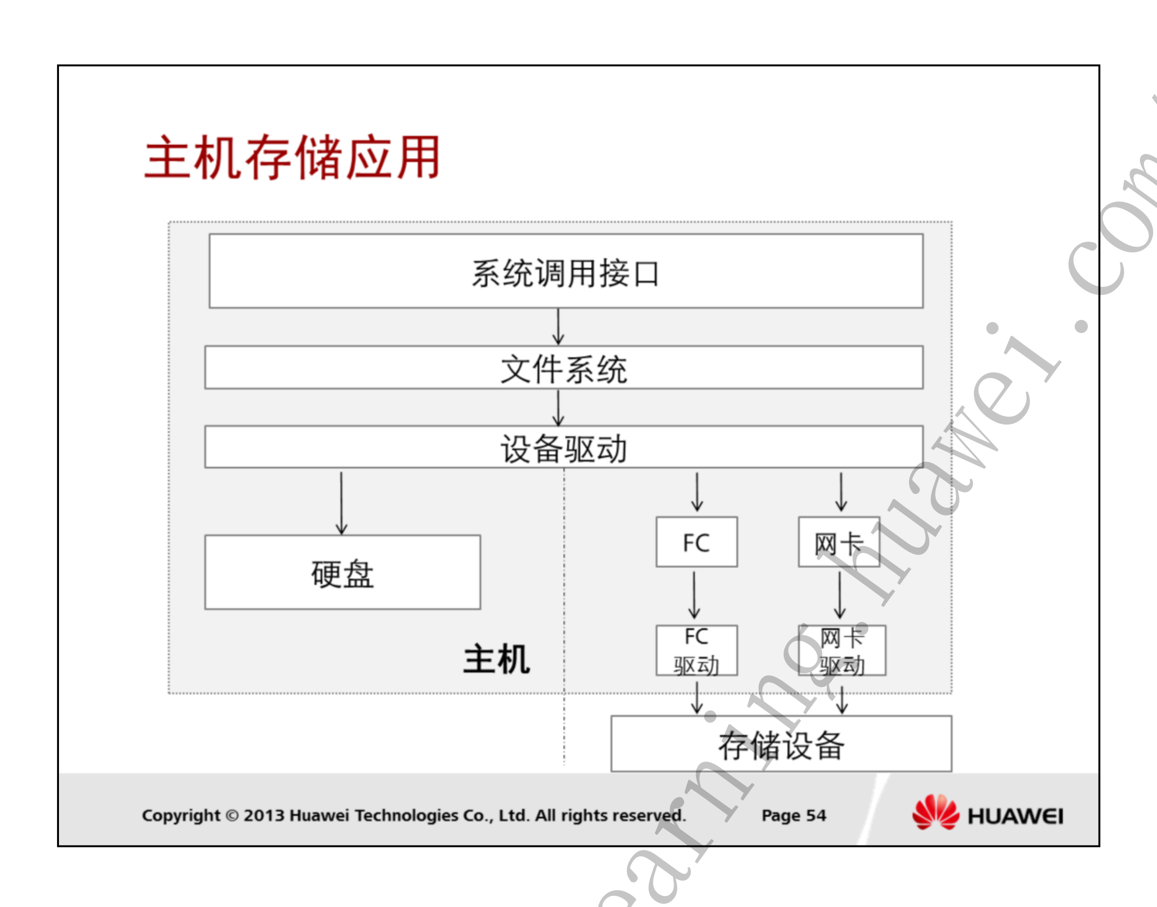
  操作系统提供外界直接或间接使用资源的管道，提供一定的安全机制来控制不同使用人员的使用权限；

- 用户界面

  提供图形化的界面，方便人机交互；

- 驱动程序

  通过驱动程序来屏蔽各厂商设备的差异，操作系统提供统一的接口来管理设备。



图的左边为主机内部的存储IO，大部分I/O开始于要访问数据的应用，应用通常不管存储的细节，而是直接调用由操作系统提供的系统调用接口，然后由文件系统为数据提供数据的逻辑地址和在磁盘上存储的物理地址的映射，再通过设备驱动层，主要是SCSI协议的操作，将数据存储到硬盘上。

图的右边为主机通过网络的存储IO，首先I/O由应用发起，然后经过操作系统，由文件系统提供数据的逻辑地址和存储的物理地址的对应关系，经由设备驱动，到达FC HBA卡或者网卡，到达存储端的FC接口或者网络接口，将数据存储到存储设备上。



目录



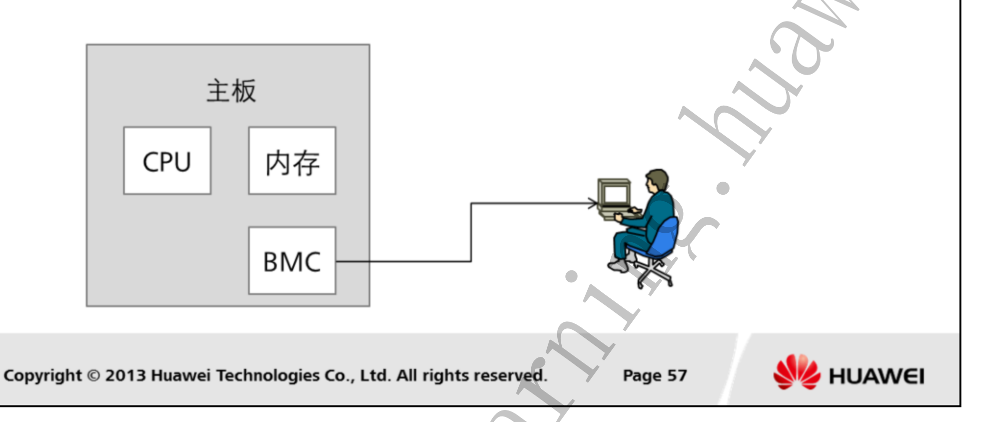


一般的主机管理方式为本地管理，而服务器大多数运行在专用的机房或者数据中心，为了实现对存放在非本地工作环境的服务器远程管理，开发了IPMI技术，通过IPMI技术可远程连接至服务器对服务器进行管理。

智能平台管理接口 (IPMI) 是一种开放标准的硬件管理接口规格，定义了嵌入式管理子系统进行通信的特定方法。IPMI 信息通过基板管理控制器 (BMC) 进行交流。使用低级硬件智能管理而不使用操作系统进行管理。

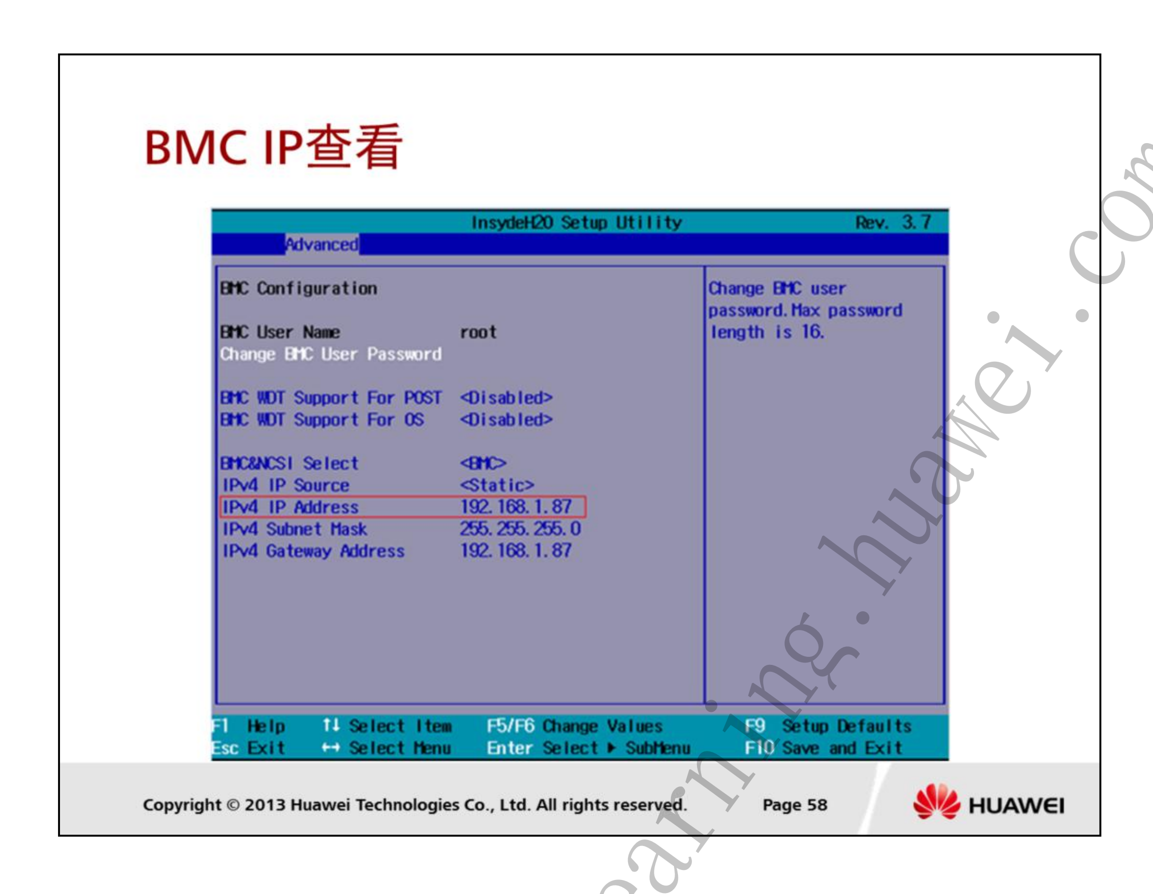
基于 Intel结构的企业系统中所使用的外围设备采用的一种工业标准，该标准由英特尔、惠普、NEC、美国戴尔电脑和SuperMicro等公司制定。用户可以利用IPMI监视服务器的物理健康特征，如温度、电压、风扇工作状态、电源状态等。而且更为重要的是IPMI是一个开放的免费标准，用户无需为使用该标准而支付额外的费用。

- 1998年Intel、DELL、HP及NEC共同提出IPMI规格，可以透过网路远端控制温度、电压
- 2001年IPMI从1.0版改版至1.5版，新增 PCI Management Bus等功能。
- 2004年Intel发表了IPMI 2.0的规格，能够向下相容IPMI 1.0及1.5的规格。新增了Console Redirection，并可以通过Port、Modem以及Lan远端管理伺服器，并加强了安全、VLAN 和刀锋伺器的支援性。



IPMI的核心是一个专用芯片/控制器(叫做服务器处理器或基板管理控制器(BMC))，其并不依赖于服务器的处理器、BIOS或操作系统来工作，是一个单独在系统内运行的无代理管理子系统。IPMI良好的自治特性便克服了以往基于操作系统的管理方式所受的限制，例如操作系统不响应或未加载的情况下其仍然可以进行开关机、信息提取等操作。

在工作时，所有的IPMI功能都是向BMC发送命令来完成的，命令使用IPMI规范中规定的指令，BMC接收并在系统事件日志中记录事件消息，维护描述系统中传感器情况的传感器数据记录。在需要远程访问系统时，IPMI新的LAN上串行(SOL)特性很有用。SOL改变IPMI会话过程中本地串口传送方向，从而提供对紧急管理服务、Windows专用管理控制台或Linux串行控制台的远程访问。BMC通过在LAN上改变传送给串行端口的信息的方向来做到这点，提供了一种与厂商无关的远程查看启动、操作系统加载器或紧急管理控制台来诊断和维修故障的标准方式。



- 在不知道BMC 管理IP的情况下，我们可以进入BIOS查看，不同的服务器主板位置可能会不一样，一般在以下位置。
    - 选择“Advanced > IPMI BMC Configuration”，按“Enter”。进入“IPMI BMC Configuration”界面。
    - 选择“BMC Configuration”，按“Enter”。进入“BMC Configuration”界面，显示BMC IP信息。



在浏览器中输入服务器的管理IP，会弹出如图的登录界面，输入用户名，密码可对服务器进行管理。默认的用户名root,默认的密码root。



可通过虚拟媒体远程安装服务器操作系统或者其它软件的安装，对于机房位置离维护人员较远的这种场景非常有用。



- 硬盘

  表示强制从硬盘启动系统。

- 光驱

  表示强制从CD/DVD启动系统。

- 软驱/可拔插移动设备

  表示强制从软驱或可拔插移动设备启动系统。

- PXE

  表示强制从PXE启动系统。

- 未配置

  表示不进行强制设置，按BIOS默认方式启动系统。



- 在“管理网口”界面中，可以进行如下配置：
  - 使能/去使能管理网口。
  - 设置管理网口的IPv4地址获取方式、IPv4地址、子网掩码、默认网关。
  - 设置管理网口的IPv6地址获取方式、IPv6地址、前缀长度、默认网关。

远程控制服务器上下电

电源管理 >> 电源控制

电源控制

上下电控制	
设备状态：下电	
上电	○
下电	○
重启	○
安全重启	○
NMI	○
	确定

 Page 63 HUAWEI

在服务器操作系统无响应的的时候，可通过远程上下电控制重新强制重启服务器。上下电控制方式包括：上电、下电、重启、安全重启、USB 复位。

- 上电

  表示对设备上电。

- 下电

  表示对设备下电。

- 重启

  表示对设备进行冷复位。

- 安全重启

  表示对设备进行软重启，即有注销的过程。

- NMI（Non-maskable Interrupt）

  表示触发业务系统产生一个不可屏蔽中断。

思考题

1. 硬盘的主要构件有哪些?
2. 固态硬盘有哪些优势?
3. 常见的文件系统有哪些?





总结

- IT基础设施与数据中心
- 存储的应用环境
- 机械硬盘与SSD硬盘结构及实现原理
- 主机应用环境





附：华为服务器产品及解决方案全景图
参考架构及解决方案
企业邮箱解决方案
HPC解决方案
数据分析解决方案
虚拟化整合解决方案
小型机迁移解决方案
数据库参考架构
SAP HANA参考架构
高密服务器
2U 2/4 Nodes X6000
44U 80 Nodes X8000
XH310
XH311
XH320
XH621
刀片服务器
刀片服务器 E6000
BH620
BH622
BH640
融合架构刀片服务器 E9000
CH121
CH220
CH221
CH222
CH240
机架服务器
1U 2S RH1288
2U 2S RH2285
2U 2S RH2288
2U 4S RH2485
2U 4S RH2488
4U 4S RH5885
8U 8S RH5885
Copyright © 2013 Huawei Technologies Co., Ltd. All rights reserved.
Page 66
HUAWEI



多选题

1、存储在数据中心的功能有（ ）

A. 提供海量存储

B. 提供集中存储

C. 提供快速I/O响应

D. 提供计算资源

2、根据服务器的形态划分，可将服务器分为（ ）

A. 机架服务器

B. 塔式服务器

C. 刀片服务器

D. 部门级服务器

答案

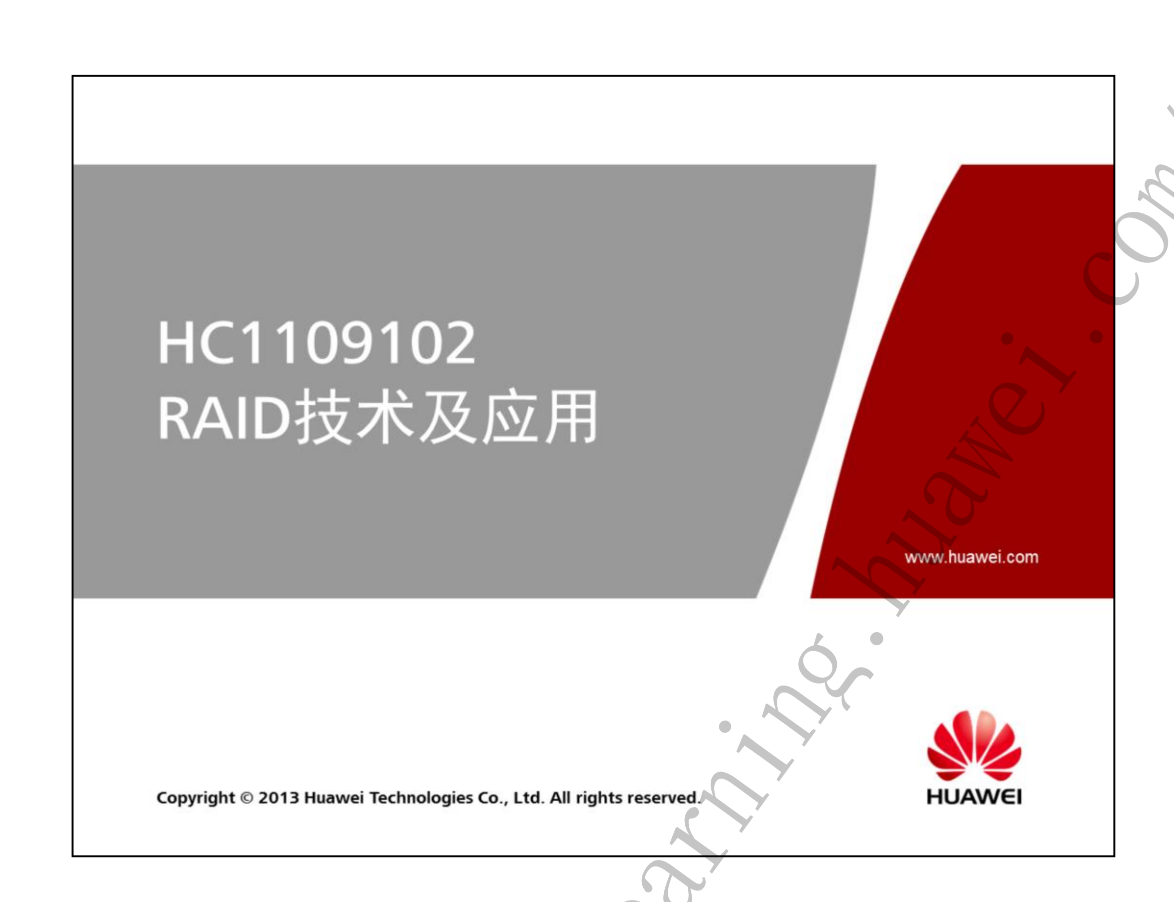
1、ABC

2、ABC


Thank you
www.huawei.com
Copyright © 2013 Huawei Technologies Co., Ltd. All rights reserved.
Page 68
HUAWEI

HC1109102 RAID 技术及应用




HC1109102
RAID技术及应用
www.huawei.com
Copyright © 2013 Huawei Technologies Co., Ltd. All rights reserved.
HUAWEI



目标

- 学习完本课程后，您将能够：
  - 了解RAID分类及原理
  - 掌握各种RAID特点
  - 掌握RAID在数据保护方面所采用的技术
  - 了解RAID2.0+技术







RAID概念与实现方式

- RAID（Redundant Array of Independent Disks）：独立冗余磁盘阵列，简称磁盘阵列。







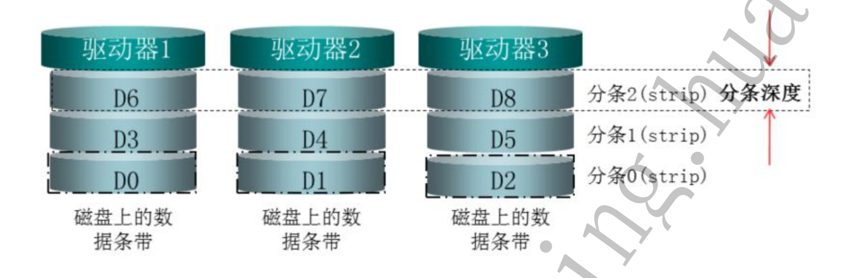
- RAID 的主要实现方式分为硬件RAID 方式和软件RAID 方式
  - 硬件RAID
  - 软件RAID

 Page 4 

- RAID技术的主要功能：
  - 通过对磁盘上的数据进行条带化，实现对数据成块存取，减少磁盘的机械寻道时间，提高了数据存取速度。
  - 通过对一阵列中的几块磁盘同时读取（并行访问），减少了磁盘的机械寻道时间，提高了数据存取速度。
  - 通过镜像或者存储奇偶校验信息的方式，实现了对数据的冗余保护。
- 目前RAID的实现方式分为硬件RAID方式和软件RAID方式。
  - 基于硬件的RAID是采用集成了处理器的RAID适配卡（简称RAID卡）来实现的。它拥有自己的控制处理器、I/O处理芯片和存储器，减少对主机CPU运算时间的占用，提高数据传输速度。 RAID控制器负责路由、缓冲以及主机和磁盘阵列之间的数据流管理。
  - 基于软件的RAID功能的实现完全依赖于主机的CPU ，没有额外的处理器和I/O芯片，所以低速CPU很难满足这个需求。软件RAID又分为基于驱动程序和基于OS两种类型。软件RAID需要占用CPU处理周期，并且依赖于操作系统，专业的企业级存储设备一般采用软件RAID的实现方式。

RAID的数据组织方式

- 条带：磁盘中单个或者多个连续的扇区构成一个条带。它是组成分条的元素。
- 分条：同一磁盘阵列中的多个磁盘驱动器上的相同“位置”（或者说是相同编号）的条带。



 Page 5



- 分条宽度
  - 指在一个分条中数据成员盘的个数；
- 分条深度
  - 指一个条带的容量大小。

RAID校验方式

- XOR校验的算法——相同为假，相异为真：

  0⊕0= 0； 0⊕1= 1； 1⊕0= 1； 1⊕1= 0；

异或校验冗余备份

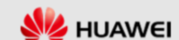
Page 6

HUAWEI

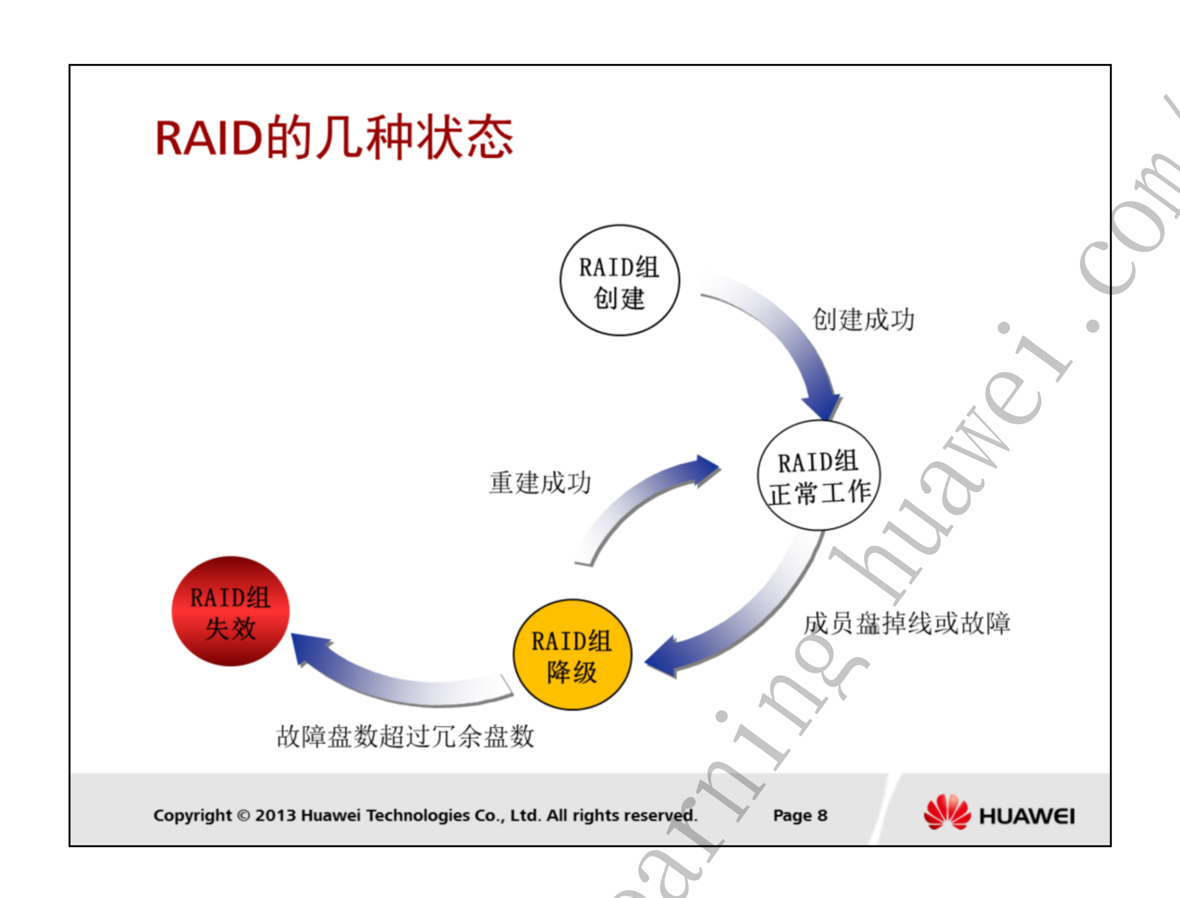
- XOR的逆运算仍为XOR：
- 如果A为1，B为0，则校验值P为1：A(1) ⊕ B(0)= P(1)
- 则有逆运算：B(0) ⊕ P(1)=A(1)；A(1) ⊕ P(1)=B(0)；

创建RAID组成员盘要求

- 硬盘容量大小　　相同
- 硬盘接口类型　　相同
- 硬盘速率要求　　相同

Page 7


- 在创建RAID组的时候，要求尽量使用同一厂家的同一型号硬盘，这样在硬盘的容量，接口，速率等方面保持一致，避免木桶效应，否则创建的RAID组工作的时候按照最低配置的硬盘的容量，性能工作，造成资源的浪费。

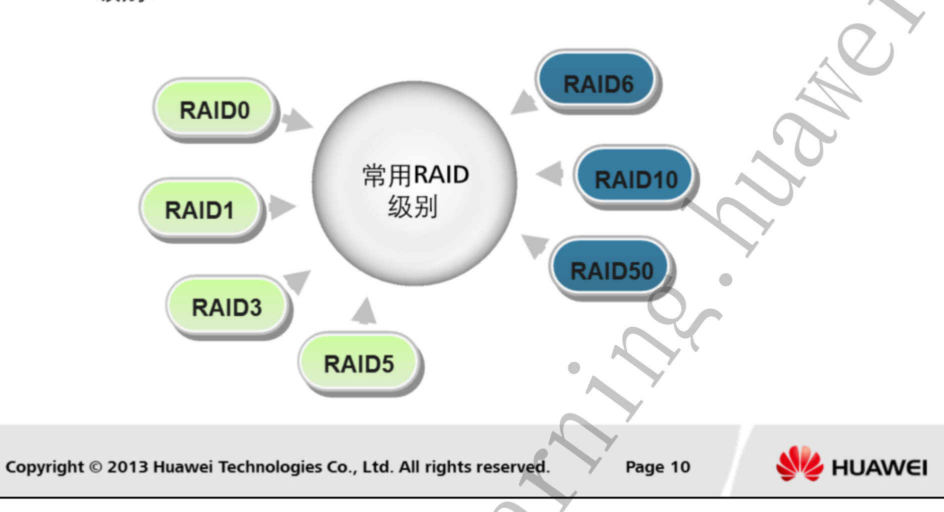


- RAID组变为降级状态后，在重建的过程中，如果还有别的成员盘出现故障，故障的成员盘的个数超过了阵列的冗余磁盘的个数，整个RAID组将为变为失效状态，此时原RAID组中的数据将会无法读取。

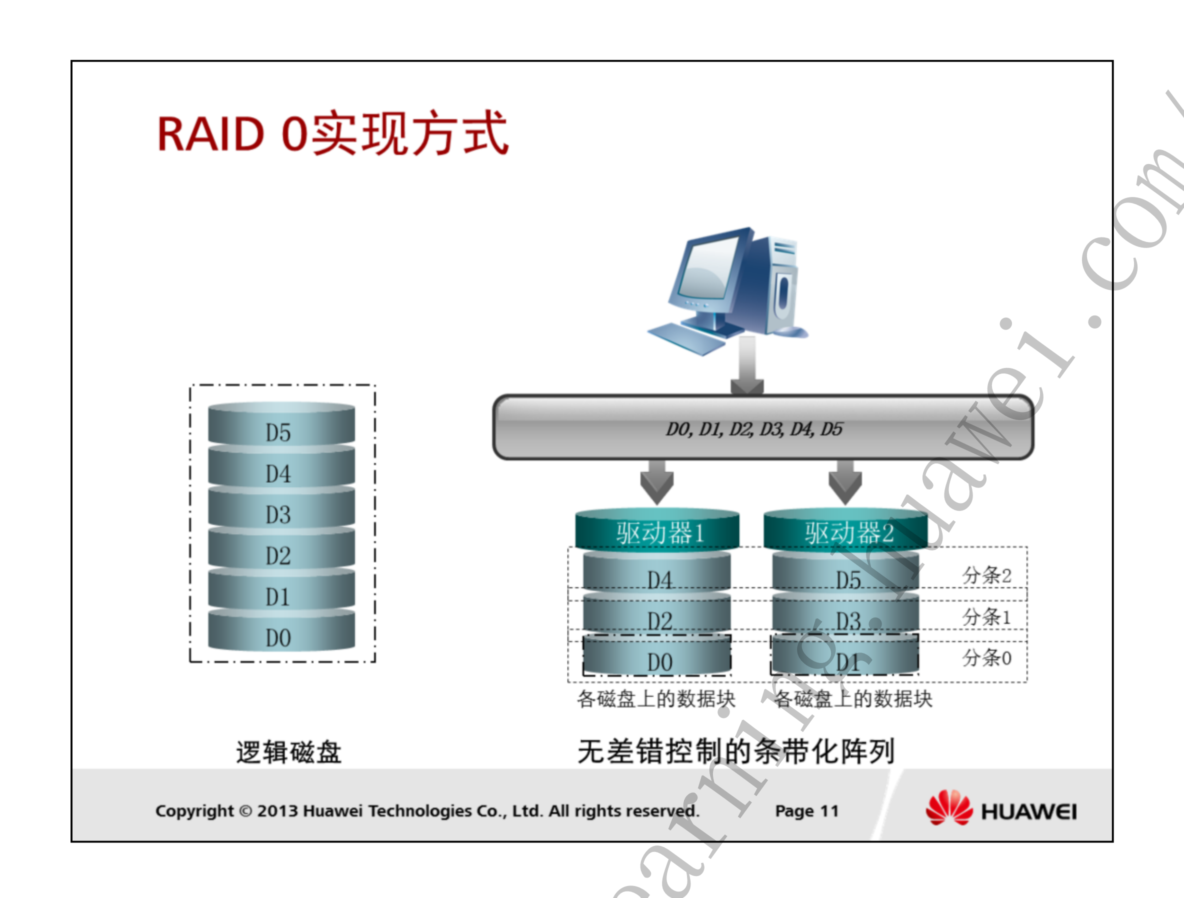


目录

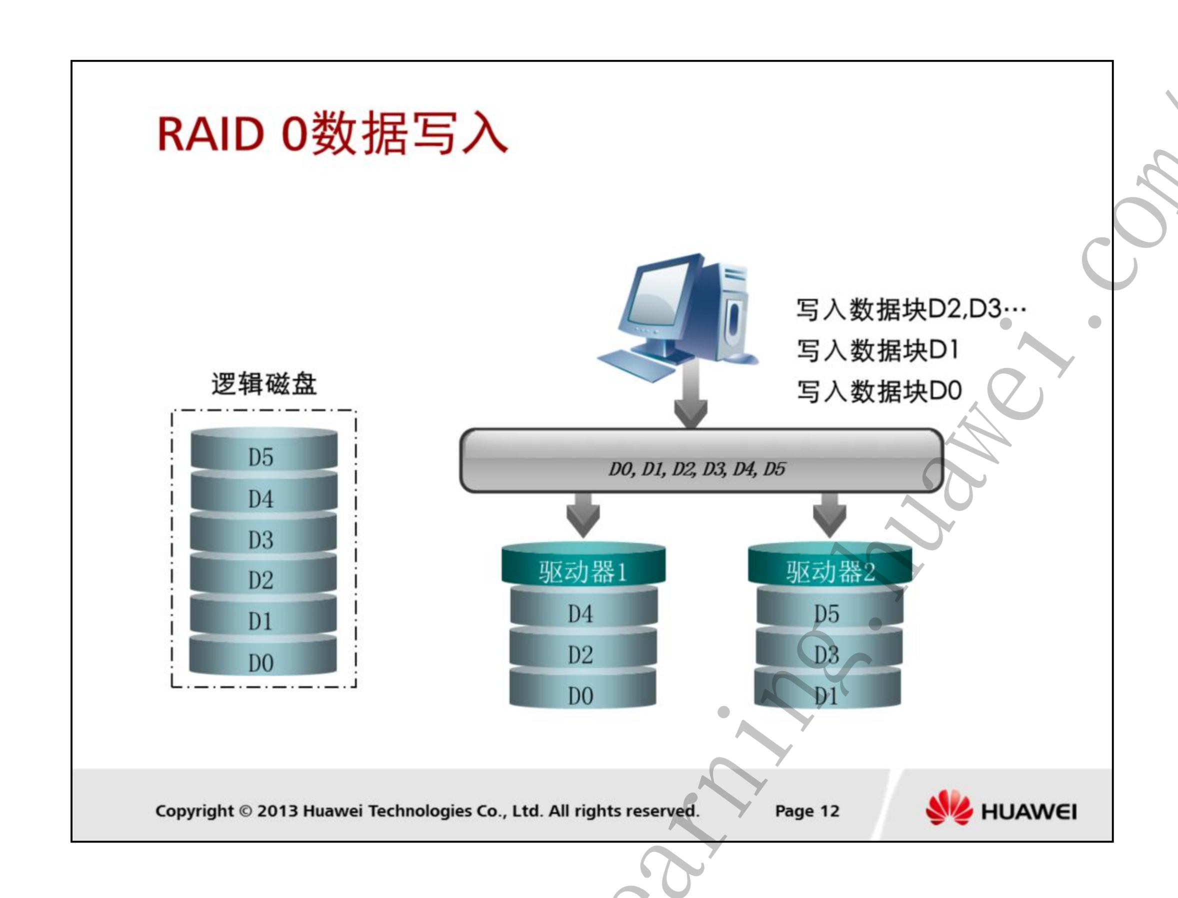




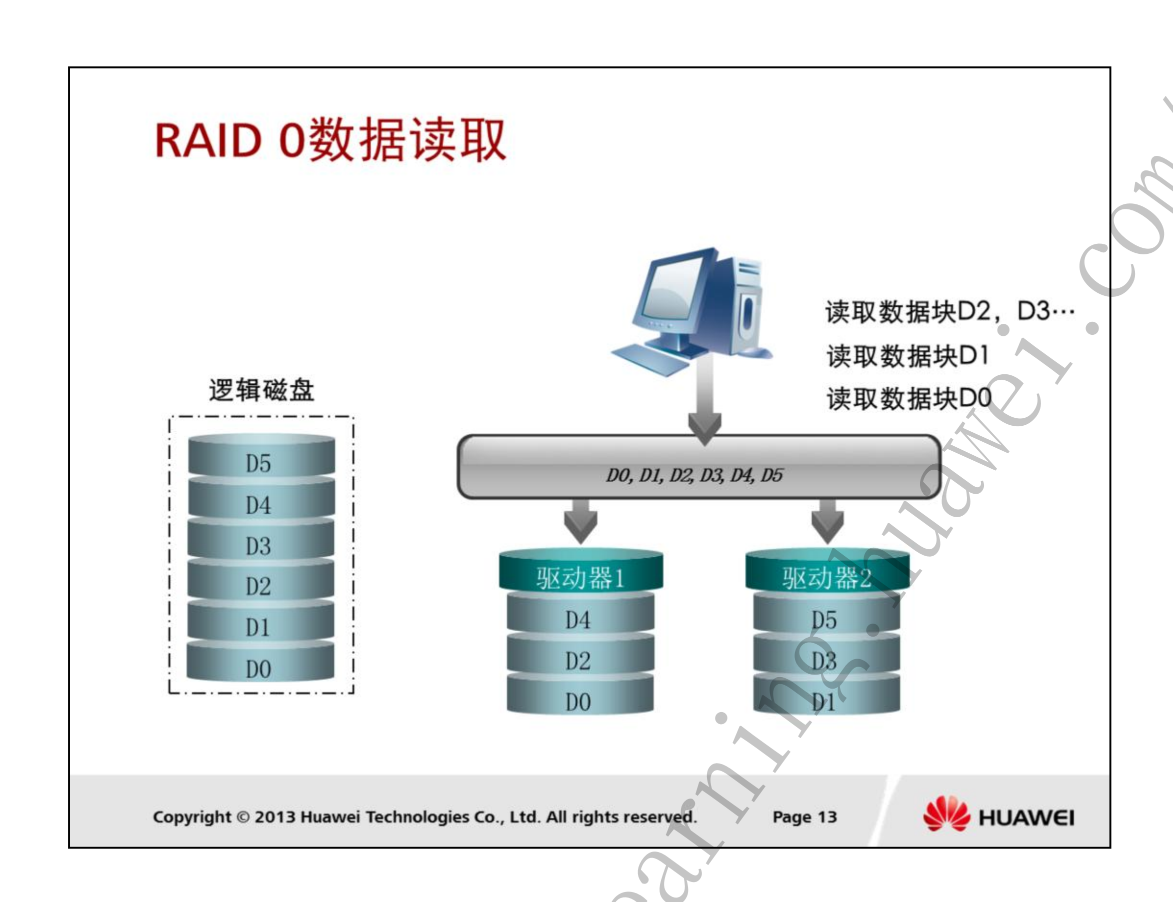
- RAID技术的优势是:
  - 通过把多个磁盘组织在一起，作为一个逻辑卷提供磁盘跨越功能。
  - 通过把数据分成多个数据块，并行写入/读出多个磁盘，以提高访问磁盘的速度。
  - 通过镜像或校验操作，提供容错能力。



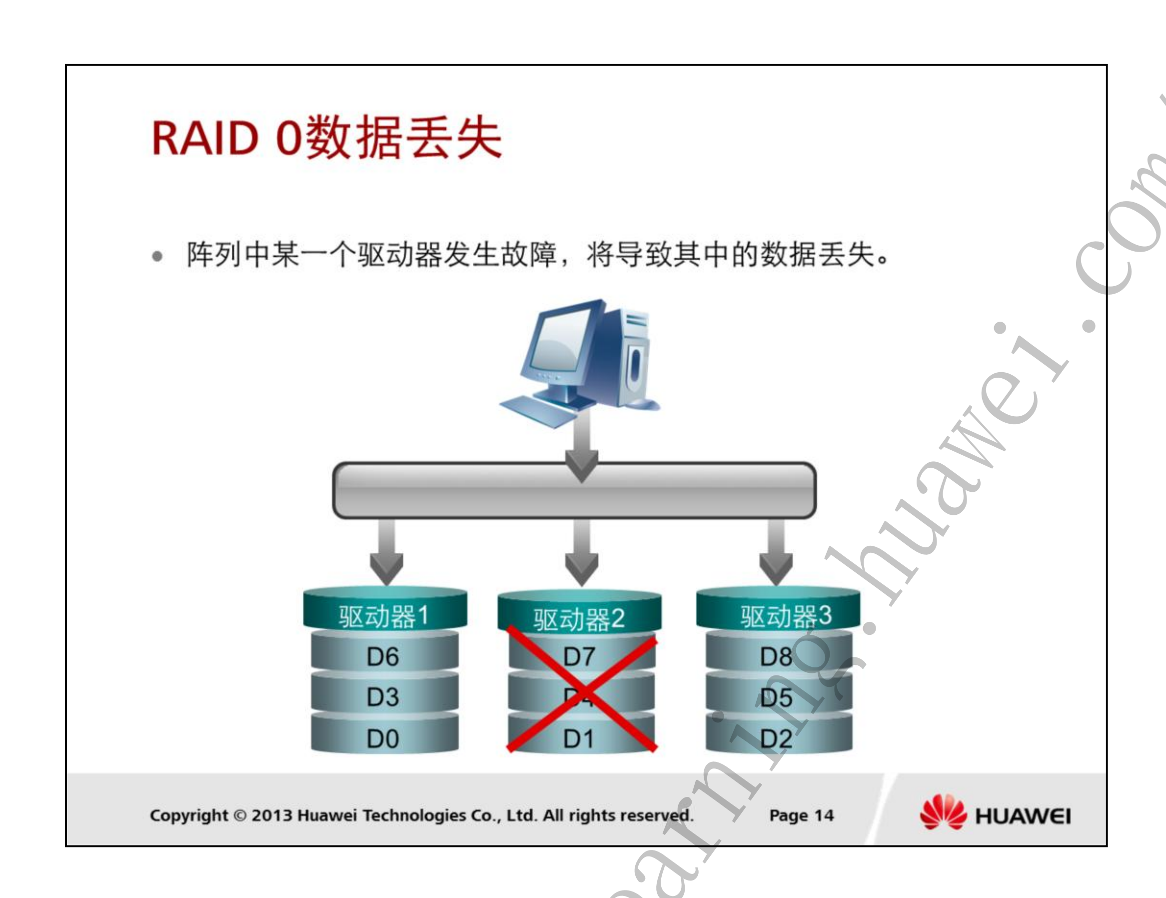
- RAID 0又称为Stripe或Striping，它代表了所有RAID级别中最高的存储性能。RAID 0使用“分条”（stripe）技术把数据分布到各个磁盘上，RAID 0至少使用两个磁盘驱动器，并将数据分成从512字节到数兆字节（一般是512Byte的整数倍）的若干块，这些数据块可以并行写到不同的磁盘中。第1块数据被写到驱动器1中，第2块数据被写到驱动器2中，如此类推，当系统到达阵列中的最后一个磁盘时，就重新回到驱动器1的下一分条进行写操作，分割数据将I/O负载平均分配到所有的驱动器。



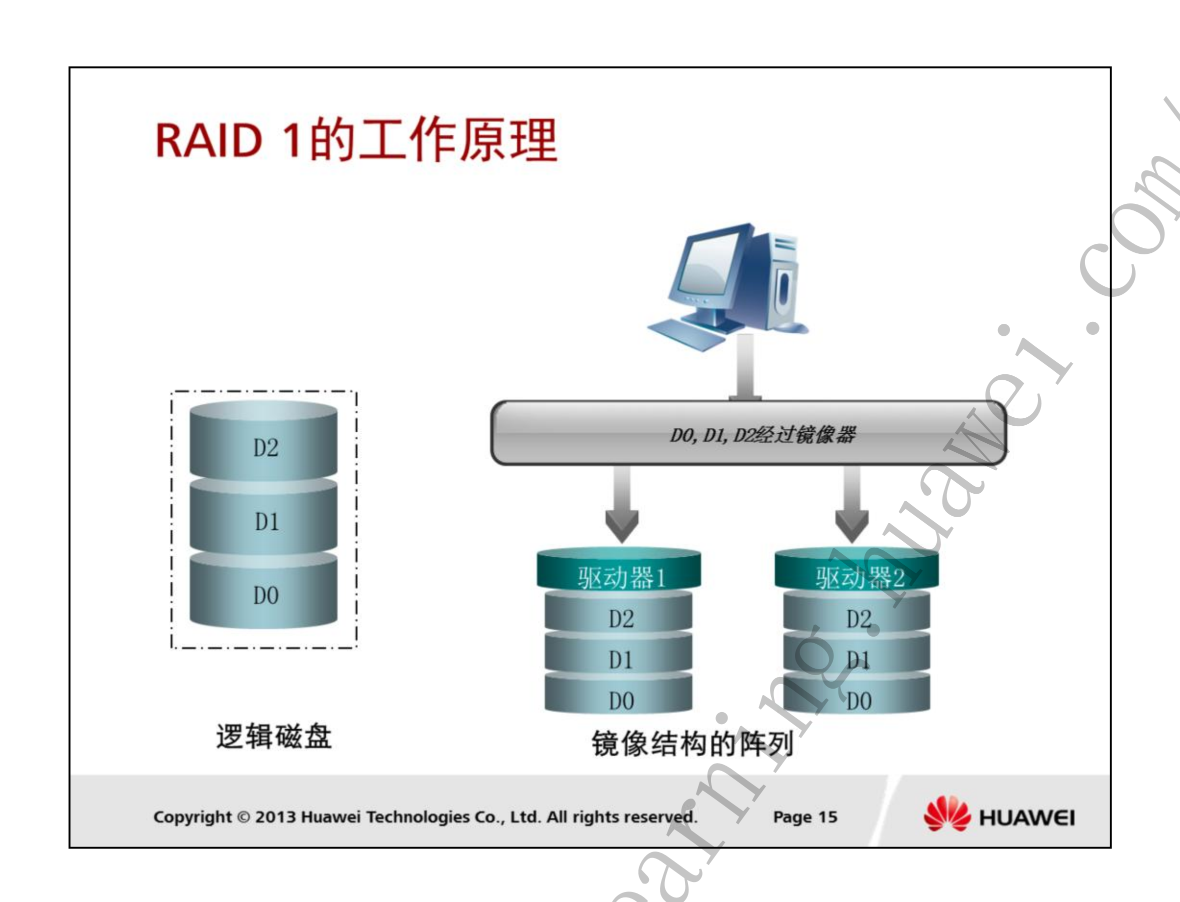
- RAID 0的数据写入是以分条形式将数据均匀分布到RAID 组的各个硬盘中。RAID 0的数据是按照分条进行写入的，即一个分条的所有分块写满后，再开始在下一个分条上进行数据写入。如上图，现在有数据D0 ，D1 ，D2 ，D3 ，D4 ，D5需要在RAID 0中进行写入，首先将第一个数据D0写入第一块硬盘位于第一个分条的块，将第二个数据D1写入第二块硬盘位于第一个分条的块，至此，第一个分条的各个块写满了数据，当有数据D2需要写入时，就要对下一个分条进行写入，将数据D2写入第一块硬盘位于第二个分条的块中… 数据块D3，D4，D5的写入同理。写满一个分条的所有块再开始在下一个分条中进行写入。



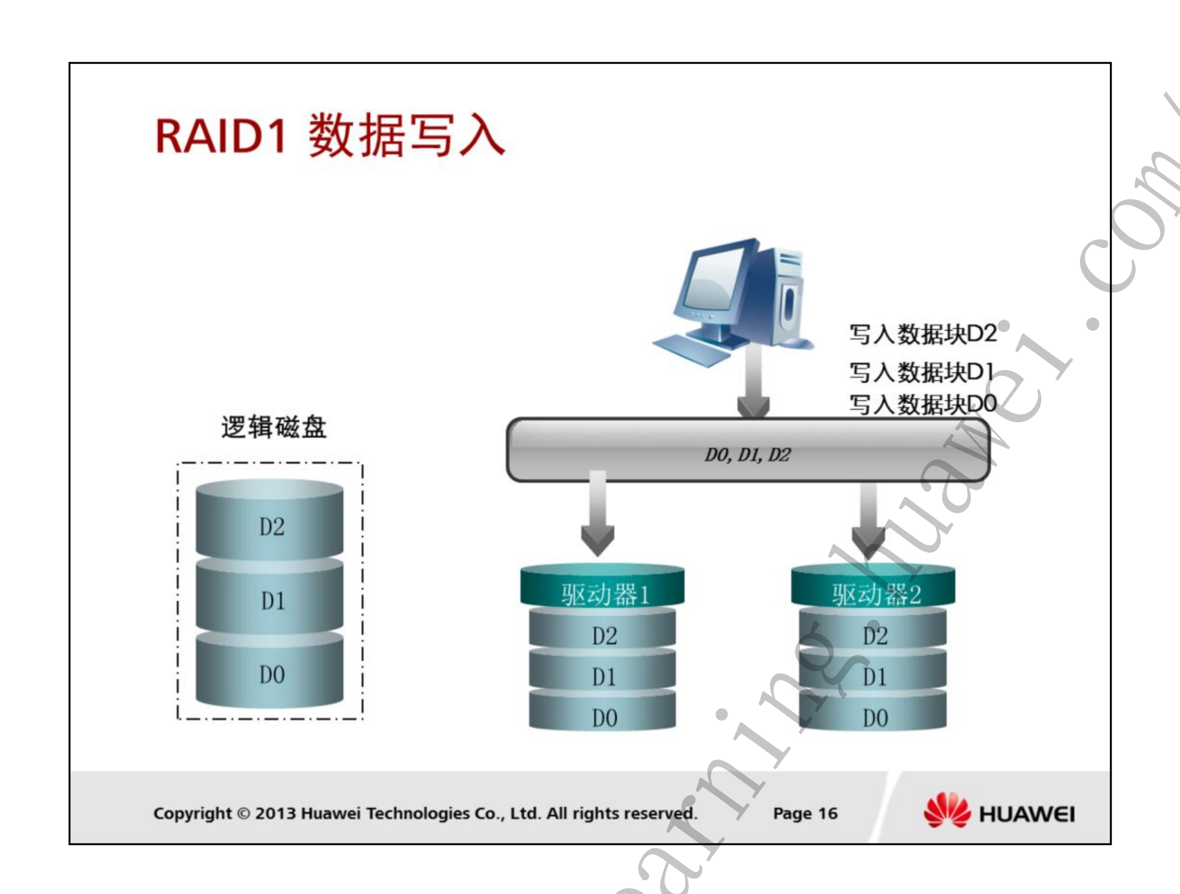
- RAID 0在收到数据读取指令后，就会在各个硬盘中进行搜索，看需要读取的数据块位于哪一个硬盘上，再依次对需要读取的数据进行读取。如上图，现在收到读取数据D0 ，D1 ，D2 ，D3 ，D4 ，D5的指令，首先从第一块磁盘读取数据块D0，再从第二块磁盘读取数据块D1…对各个数据块，从磁盘阵列读取后再由RAID控制器进行整合，传送给系统，至此，整个读取过程结束。



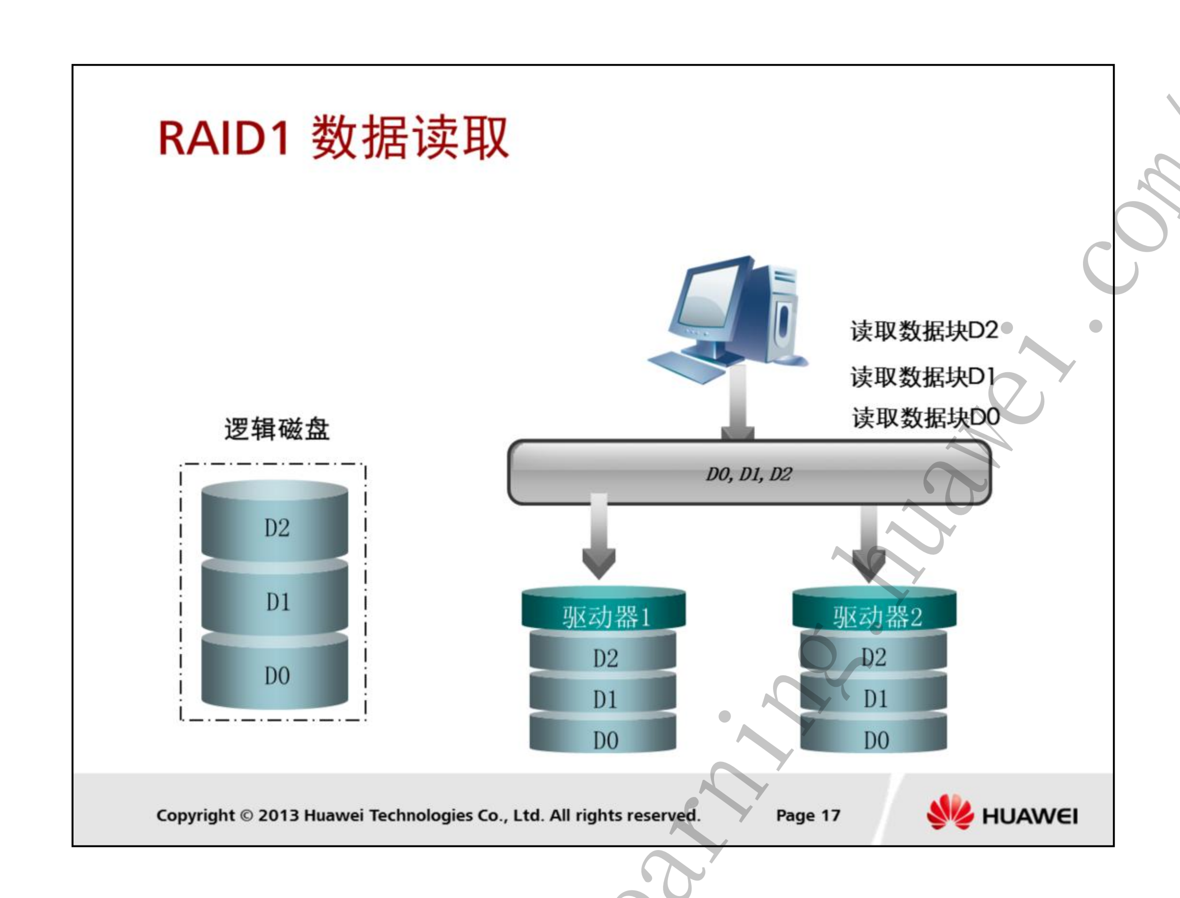
- 由于RAID 0只是将数据按一定方式组织起来，而没有在各个磁盘的数据块之间提供数据安全性保护，所以一旦阵列中有某一个驱动器发生故障，整个阵列将失效。



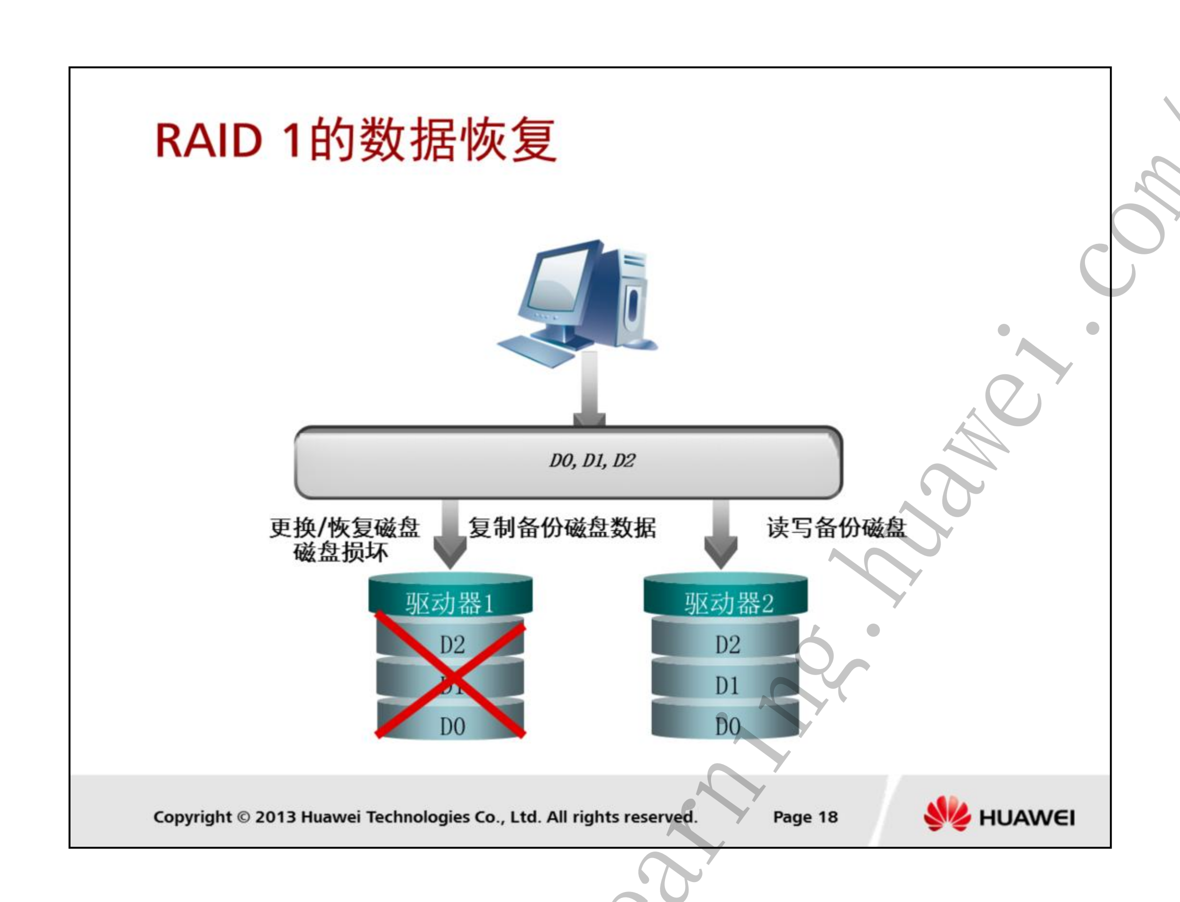
- RAID 1也被称为镜像，其目的是为了打造出一个安全性极高的RAID。RAID 1使用两组相同的磁盘系统互作镜像，速度没有提高，但是允许单个磁盘故障，数据可靠性最高。其原理为在主硬盘上存放数据的同时也在镜像硬盘上写一样的数据。当主硬盘（物理）损坏时，镜像硬盘则代替主硬盘的工作。因为有镜像硬盘做数据备份，所以RAID 1的数据安全性在所有的RAID级别上来说是最好的。



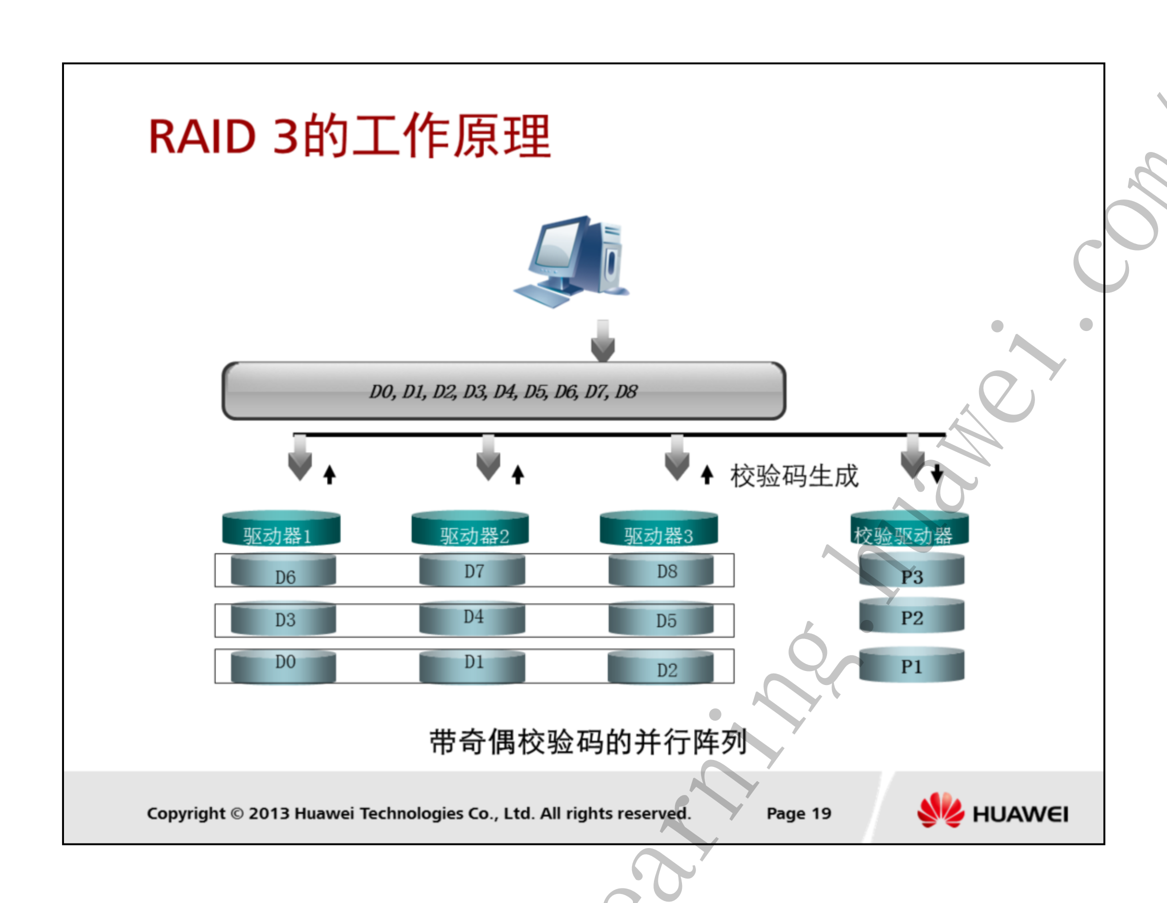
- RAID 1在进行数据写入的时候，并不是像RAID 0那样将数据分条写入所有磁盘，而是将数据分别写入成员盘，各个成员磁盘上的数据完全相同，互为镜像。如上图，需要将数据块D0，D1，D2写入RAID 1，先在两个磁盘上同时写入数据块D0，再在两个磁盘上同时写入数据块D1…



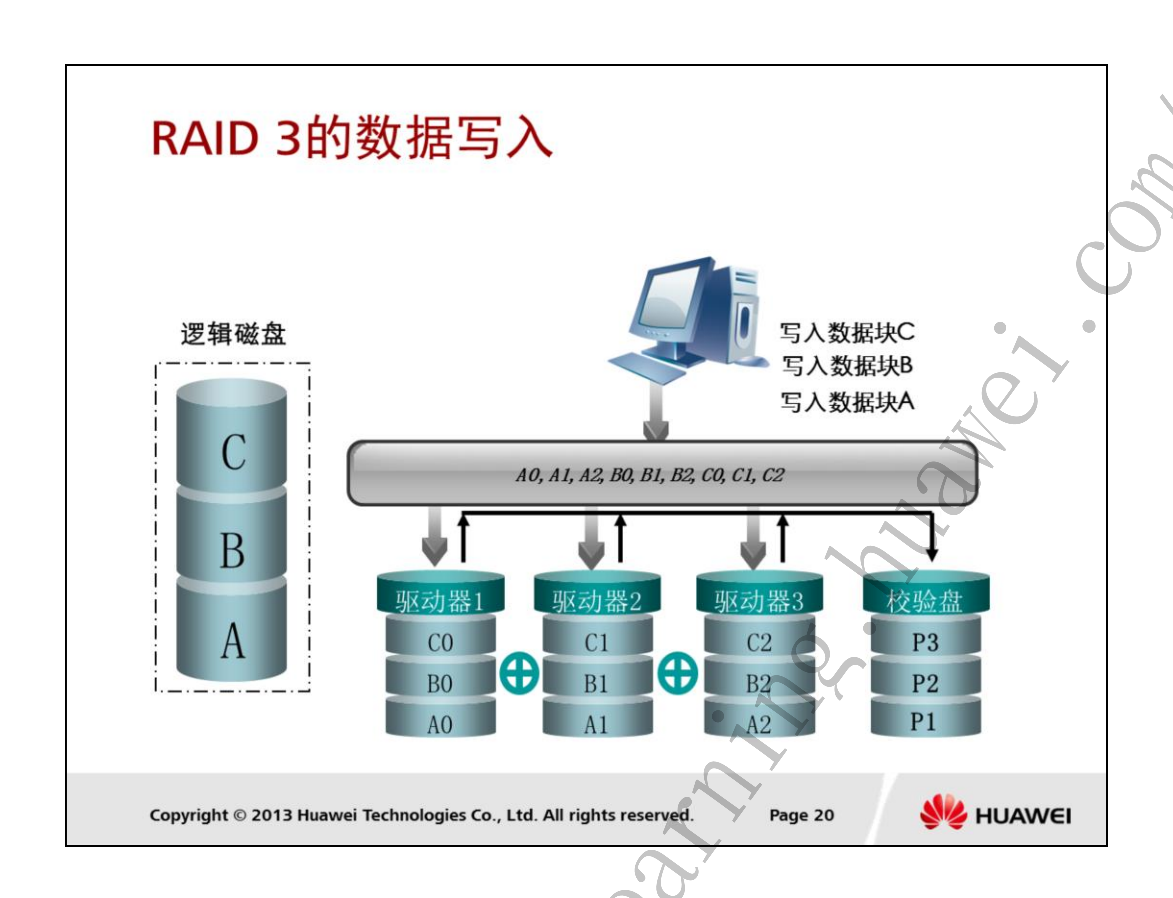
- RAID 1在进行数据读取的时候，正常情况下可以实现数据盘和镜像盘同时读取数据，提高读取性能，如果一个磁盘损坏，则IO自动到存活的盘读取数据。



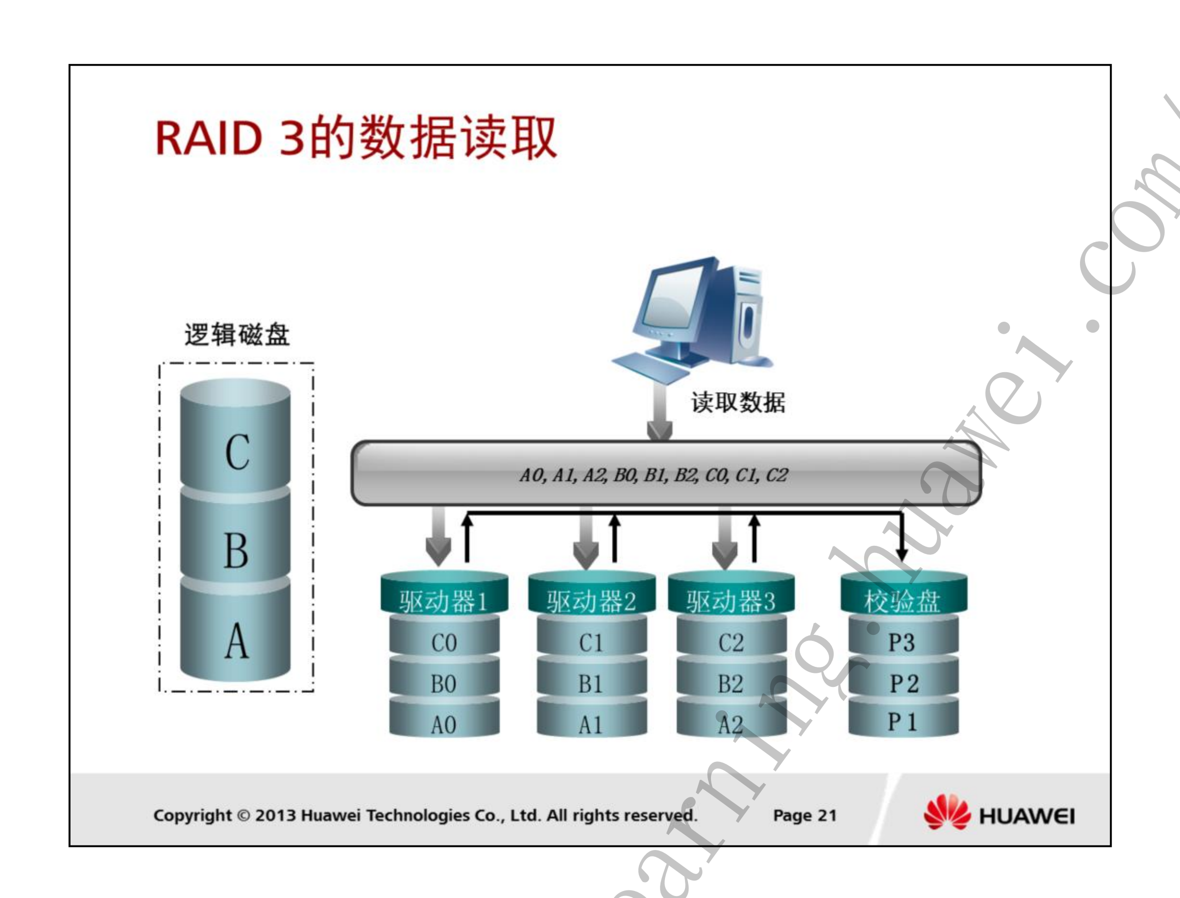
- RAID 1的成员磁盘是互为镜像的，成员磁盘的内容完全相同，这样，任何一组磁盘中的数据出现问题，都可以马上从其它成员磁盘进行镜像恢复。如上图，磁盘1损坏导致数据丢失，我们需要将故障磁盘用正常磁盘替换，再读取磁盘2的数据，将其复制到磁盘1上，从而实现了数据的恢复。



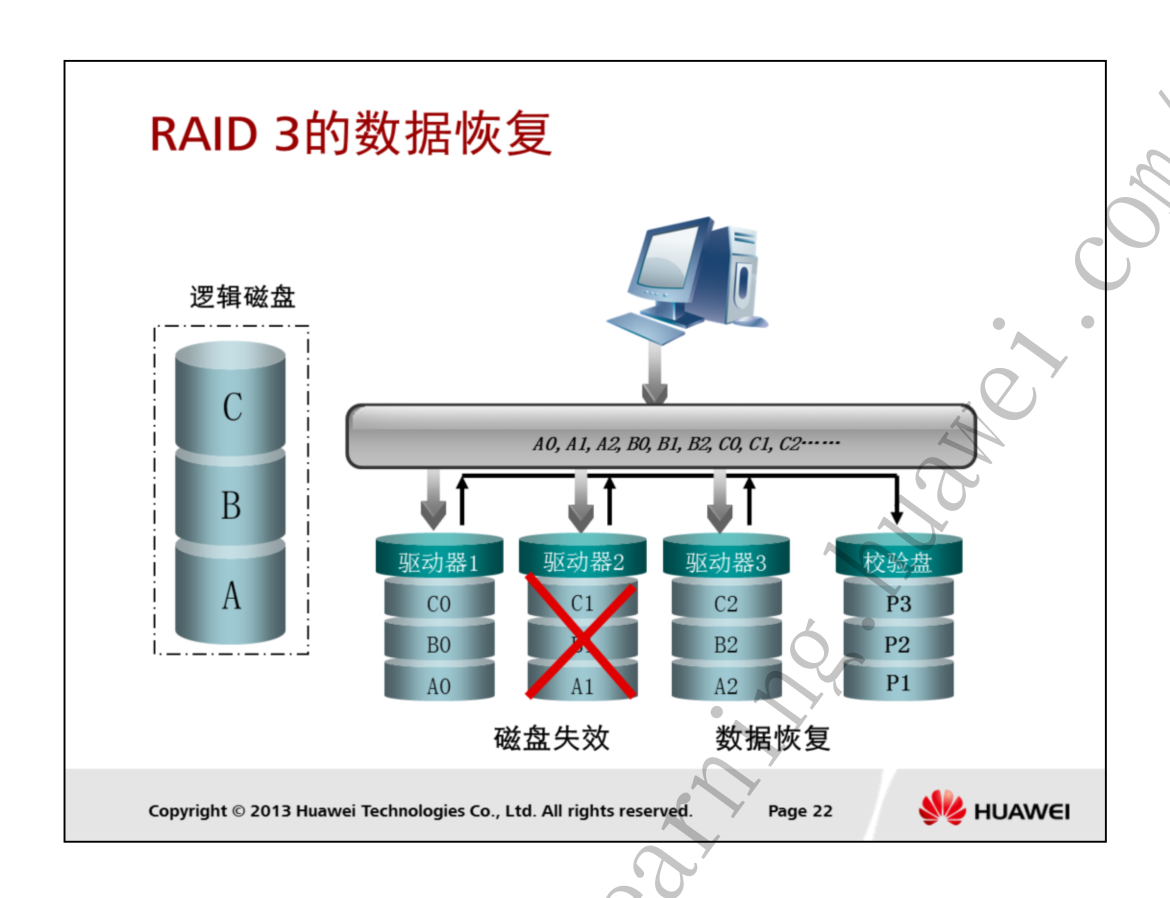
- RAID 3是带有专用奇偶位的条带化阵列，是RAID 0的一种改进模式。它也采用了类似于RAID 2模式中的奇偶校验技术，不过没有使用海明码技术而采用较为简单的异或算法。在阵列中有一个驱动器专门用来保存其它驱动器中对应分条中数据的奇偶校验信息。奇偶位是编码信息，如果某个驱动器中的数据出错或者某一个驱动器故障，可以通过对奇偶校验信息的计算来恢复出故障驱动器中的数据信息。在数据密集型环境或者是单一用户环境中，组建RAID 3对访问较长的连续记录较好。在写入数据时，RAID 3会把数据的写入操作分散到多个磁盘上进行，然而不管是向哪一个数据盘写入数据，都需要同时重写校验盘中的相关信息。因此，对于那些经常需要执行大量写入操作的应用来说，校验盘的负载将会很大，无法满足程序的运行速度，从而导致整个RAID系统性能的下降。



- RAID 3是单盘容错并行传输。即采用Stripping技术将数据分块，对这些块进行异或校验，校验数据写到最后一个硬盘上。它的特点是有一个盘为校验盘，数据以位或字节的方式存于各盘（分散记录在组内相同扇区的各个硬盘上）。当一个硬盘发生故障，除故障盘外，写操作将继续对数据盘和校验盘进行操作。



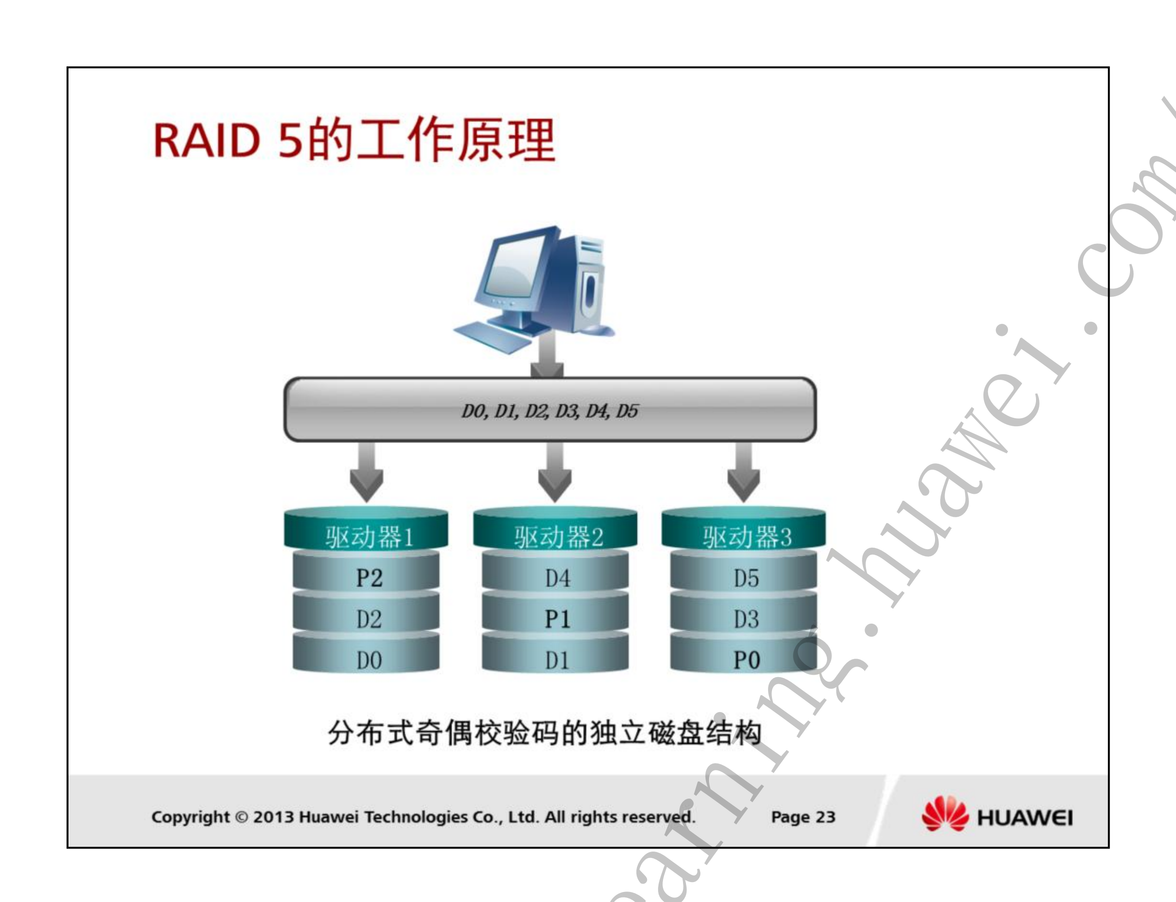
- RAID3的数据读取是按照分条来进行的。将每个磁盘的驱动器主轴马达做精确的控制，同一分条上各个磁盘上的数据位同时读取，各个驱动器得到充分利用，读性能较高。RAID 3的数据读写属于并行方式。



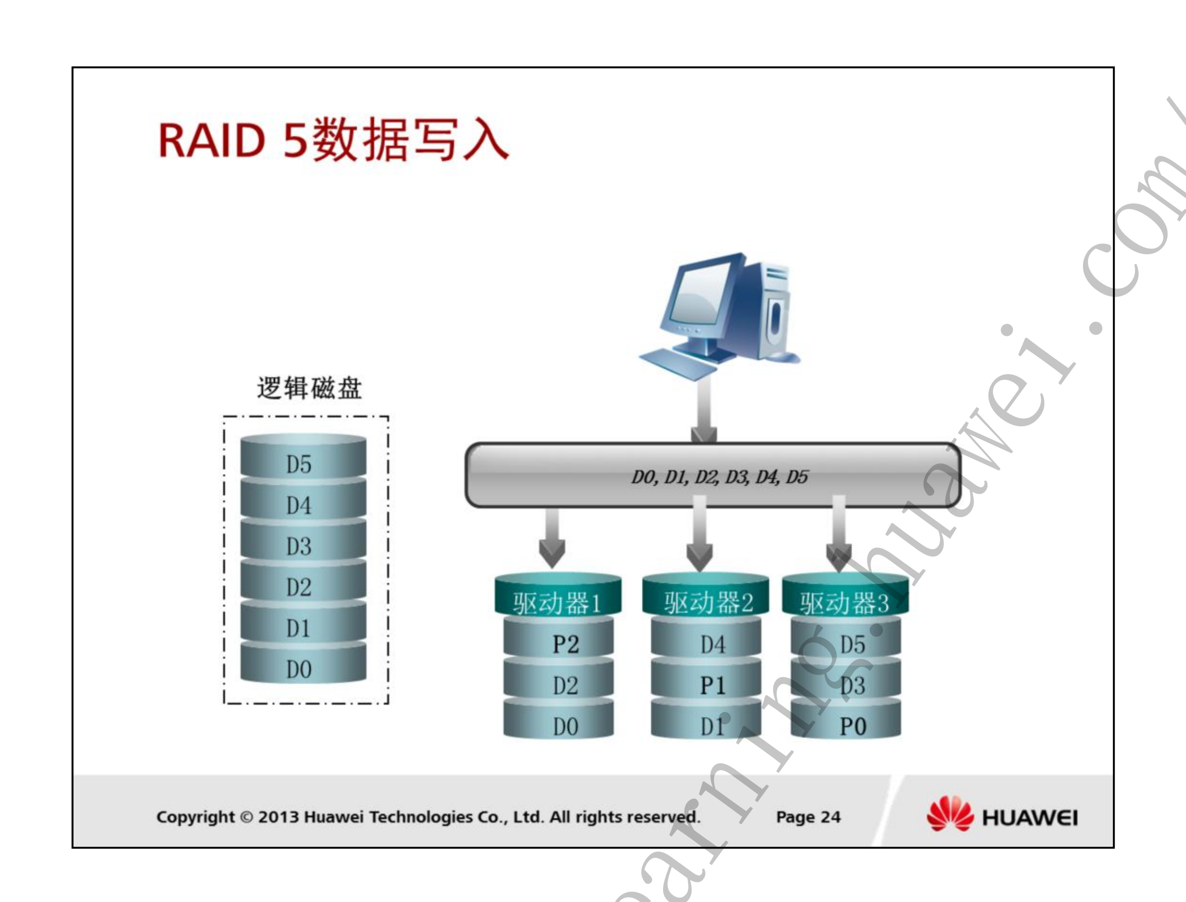
RAID 3的数据恢复是通过对剩余数据盘和校验盘的异或计算重构故障盘上应有的数据来进行的。

如上图的RAID 3磁盘结构，当磁盘2故障，其上存储的数据位A1，B1，C1丢失，我们需要经过这样一个数据恢复过程：首先恢复数据A1，根据同一分条上其它数据盘和校验盘上的数据A0，A2，P1，进行异或运算，得到应有的数据A1，再用相同方法恢复出数据B1，C1的数据，至此，磁盘2上的数据全部得到了恢复。

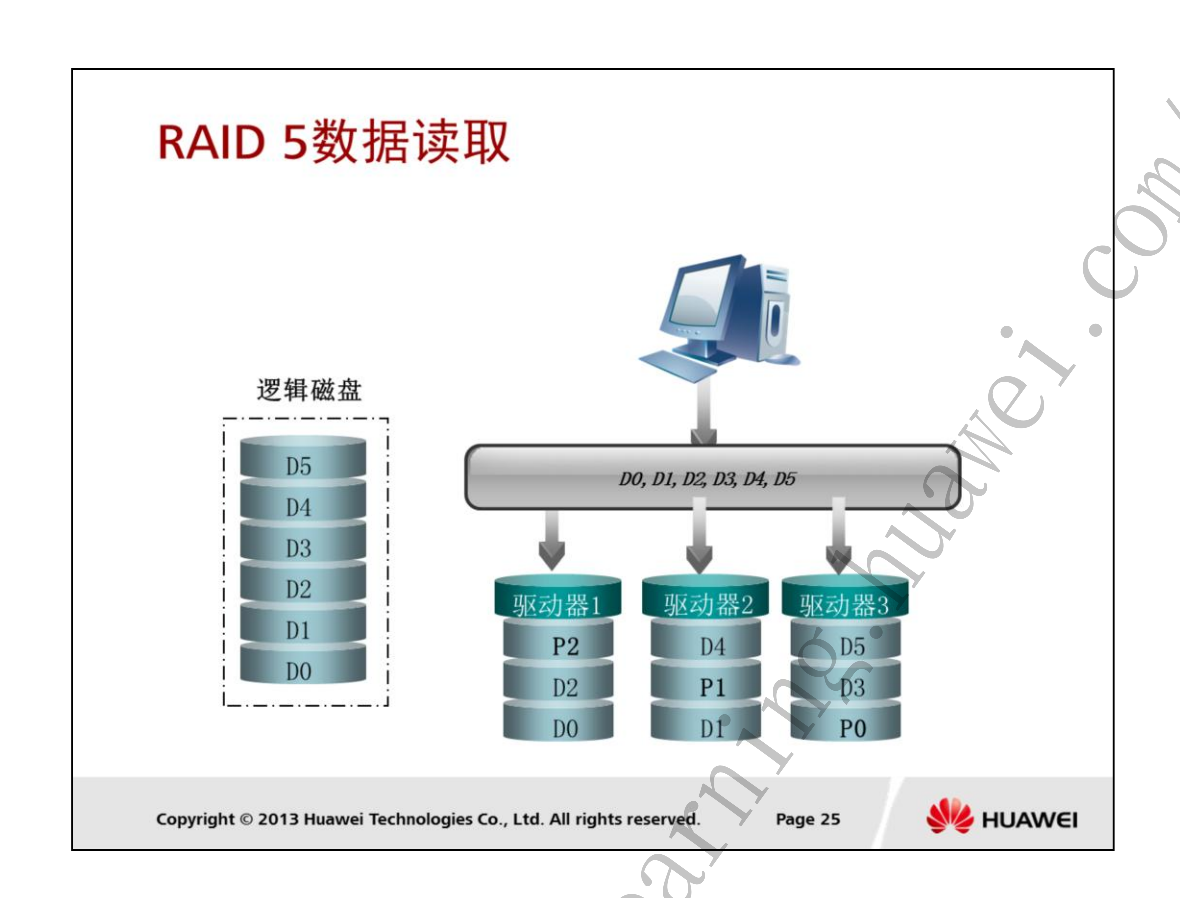
由于校验集中在一个盘，因此在数据恢复时，校验盘写压力比较大，影响性能。



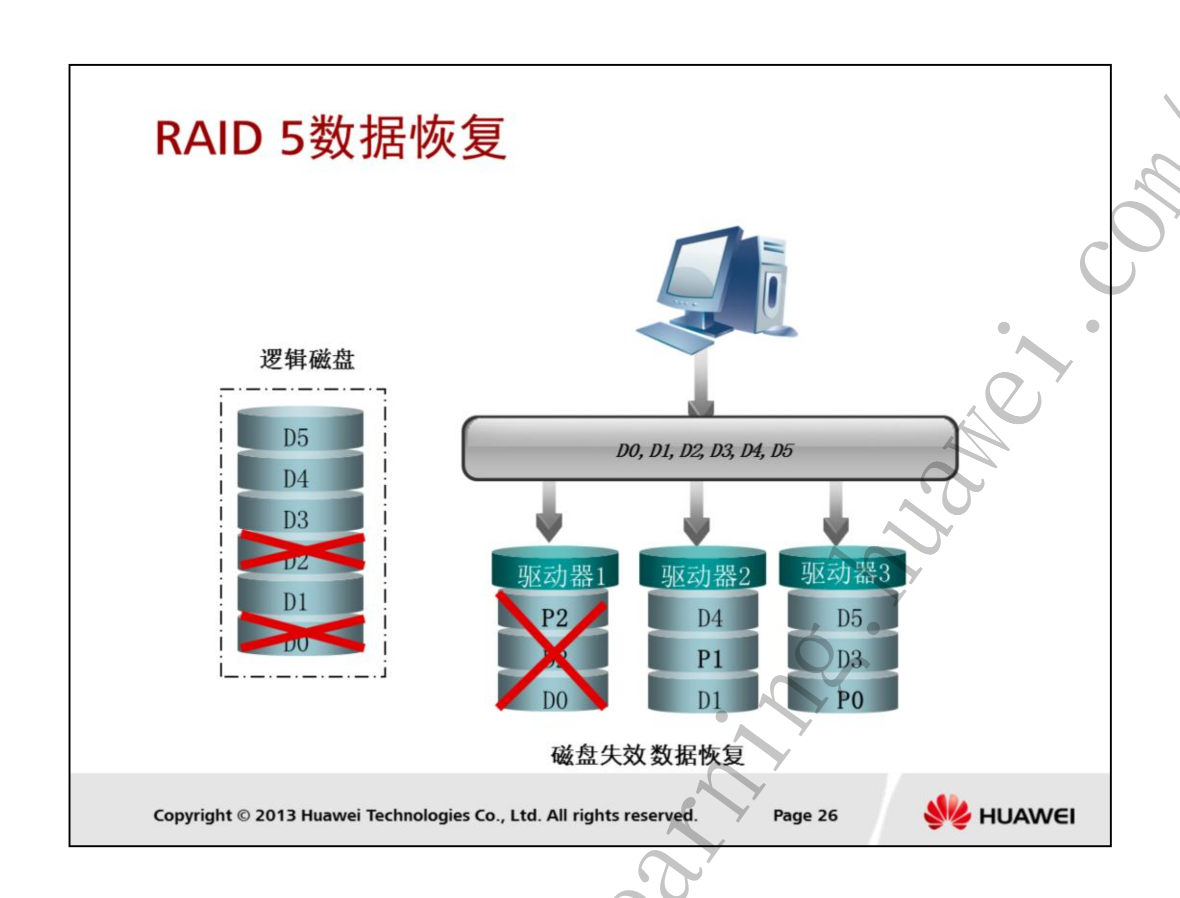
RAID5是一种旋转奇偶校验独立存取的阵列方式，它与RAID3不同的是没有固定的校验盘，而是按某种规则把奇偶校验信息均匀地分布在阵列所属的硬盘上，所以在每块硬盘上，既有数据信息也有校验信息。这一改变解决了争用校验盘的问题，能并发进行多个写操作。所以RAID 5即适用于大数据量的操作，也适用于各种事务处理，它是一种快速、大容量和容错分布合理的磁盘阵列。当有N块阵列盘时，可用容量为N-1块盘容量。 RAID 3、RAID 5中，在一块硬盘发生故障后，RAID组从ONLINE变为DEGRADED方式，直到故障盘恢复。但如果在DEGRADED状态下，又有第二块盘故障，整个RAID组的数据将丢失。



RAID 5的数据写入也是按分条进行的，各个磁盘上既存储数据块，又存储校验信息。一个分条上的数据块写入完成后，将产生的校验信息写入对应的校验磁盘中。



由于RAID 5的数据是按照数据分条存储的，在读取的时候，按照分条进行读取。



当RAID5中某一个磁盘故障，在恢复的时候，可利用其它存活成员盘数据进行异或逆运算，恢复故障盘上的数据。

RAID 6介绍

- RAID 6是带有两种校验的独立磁盘结构，采用两种奇偶校验方法，需要至少N+2(N>2)个磁盘来构成阵列，一般用在数据可靠性、可用性要求极高的应用场合 。
- 常用的RAID 6技术有RAID6 P＋Q和RAID6 DP。

 Page 27



RAID 6实际上是在RAID 5基础上为了进一步保证数据可用性和可靠性设计的一种RAID方式。与RAID 5相比除了有通常的异或校验方式外，还增加了另一种特殊的异或校验方式和该方式校验数据存放区域，因此RAID 6的数据冗余性能相当好。但是，由于增加了一个校验，所以写入的效率比RAID 5要低，而且控制系统的设计也更为复杂，第二个校验区也减少了有效存储空间。

- 目前RAID 6还没有统一的标准，各家公司的实现方式都有所不同，主要有以下两种方式：
  - RAID P+Q： 华为、HDS
  - RAID DP： NetApp

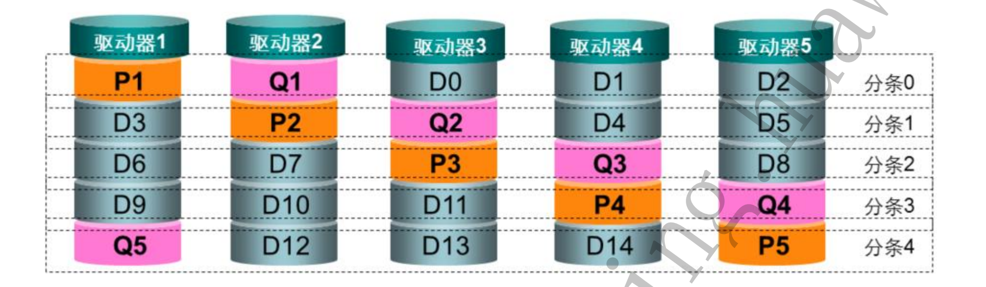
两种技术获取校验信息的方法不同，但是都能够在两块成员盘故障的情况下读取数据，数据不丢失。

RAID6 P＋Q的工作原理

- RAID6 P＋Q需要计算出两个校验数据P和Q，当有两个数据丢失时，根据P和Q恢复出丢失的数据。校验数据P和Q是由以下公式计算得来的：

$$P=D0\oplus D1 \oplus D2 \ldots\ldots$$

$$Q=(\alpha\otimes D0)\oplus(\beta\otimes D1)\oplus(\gamma\otimes D2)\ldots\ldots$$



 Page 28



在RAID6 P＋Q中，P和Q是两个相互独立的校验值，它们的计算互不影响，都是由同一分条上其它数据磁盘上的数据依据不同的算法计算而来的。

其中P值的获得是通过同一分条上除P和Q之外的其它所有数据盘上数据的简单异或运算得到。Q值的获得过程就相对复杂一些，它首先对同一分条其他磁盘上的各个数据分别进行一个变换，然后再将这些变换结果进行异或操作而得到校验盘上的数据。这个变换被称为GF变换，它是一种常用的数学变换方法，可以通过查GF变换表而得到相应的变换系数，再将各个磁盘上的数据与变换系数进行运算就得到了GF变换后的数据，这个变换过程是由RAID控制器来完成的。

以上图为例，P1是由分条0中的数据D0、D1、D2进行简单的异或运算而得到的。同理，P2是由分条1中的数据D3、D4、D5进行简单的异或运算而得到， P3是由分条2中的数据D6、D7、D8进行简单的异或运算而得到。

Q1是由分条0中的数据D0、D1、D2分别进行GF变换之后再进行异或运算而得到的。同理，Q2是由分条1中的数据D3、D4、D5分别进行GF变换之后再进行异或运算而得到，Q3是由分条2中的数据D6、D7、D8分别进行GF变换之后再进行异或运算而得到。

当某一个分条中有一块磁盘发生故障，根本不需要Q，直接用校验值P与其他正常数据进行异或运算就可以恢复出故障盘上面的数据，数据恢复比较方便。当分条中有两块磁盘发生故障，如果其中包含Q所在的磁盘，则可以先恢复出数据盘上面的数据，再恢复出校验盘Q上的校验值；如果故障盘不包含Q所在的盘，则可以将两个校验公式作为方程组，从而可以恢复出两个故障盘上面的数据。

RAID6 DP的工作原理

- DP – Double Parity，就是在RAID4所使用的一个行XOR校验磁盘的基础上又增加了一个磁盘用于存放斜向的XOR校验信息
- 横向校验盘中P0—P3为各个数据盘中横向数据的校验信息

  例：P0=D0 XOR D1 XOR D2 XOR D3
- 斜向校验盘中DP0—DP3为各个数据盘及横向校验盘的斜向数据校验信息

  例：DP0=D0 XOR D5 XOR D10 XOR D15



 Page 29 

RAID6 DP 同样也有两个相互独立的校验信息块，但是与RAID6 P＋Q 不同的是，它的第二块校验信息是斜向的。横向校验信息和斜向校验信息都使用异或校验算法而得到，横向校验盘中的信息的获得方法非常简单：P0是由条带0中的数据D0、D1、D2、 D3进行简单的异或运算而得到的。同理， P1是由分条1中的数据D4、D5、D6、D7进行简单的异或运算而得到……

即P0=D0 ⊕ D1 ⊕ D2 ⊕ D3；P1=D4 ⊕ D5 ⊕ D6 ⊕ D7

斜向校验盘中校验信息的获得依然是采用数据之间的异或运算，只是数据块的选取相对复杂一些，是一个斜向的选取过程：

由分条0上面第一个磁盘上的数据D0、分条1上面第二个磁盘上的数据D5、分条2上面第三个磁盘上的数据D10、分条3上面第四个磁盘上的数据D15经过异或校验而得到校验信息DP0；由分条0上面第二个磁盘上的数据D1、分条1上面第三个磁盘上的数据D6、分条2上面第四个磁盘上的数据D11、分条3上面校验盘上面的信息P3经过异或校验而得到校验信息DP1；由分条0上面第三个磁盘上的数据D2、分条1上面第四个磁盘上的数据D7、分条2上面校验盘上面的信息P2、分条3上面第一个磁盘上的数据D12经过异或校验而得到校验信息DP2……

即DP0=D0 ⊕ D5 ⊕ D10 ⊕ D15；DP1=D1 ⊕ D6 ⊕ D11 ⊕ P3

RAID6 DP允许阵列中同时有两个磁盘失效，我们以上图为例，假设磁盘1、2故障，则数据D0、D1、D4、D5、D8、D9、D12、D13失效，其他磁盘数据和校验信息正常。我们来看一下数据恢复是怎样一个过程：首先根据DP2和斜向校验恢复出D12， D12 =D2 ⊕ D7 ⊕ P2 ⊕ DP2，然后利用P3和横向校验恢复出D13， D13 =D12 ⊕ D14 ⊕ D15 ⊕ P3 ；

RAID6 DP的工作原理

- DP – Double Parity，就是在RAID4所使用的一个行XOR校验磁盘的基础上又增加了一个磁盘用于存放斜向的XOR校验信息
- 横向校验盘中P0—P3为各个数据盘中横向数据的校验信息

  例：P0=D0 XOR D1 XOR D2 XOR D3
- 斜向校验盘中DP0—DP3为各个数据盘及横向校验盘的斜向数据校验信息

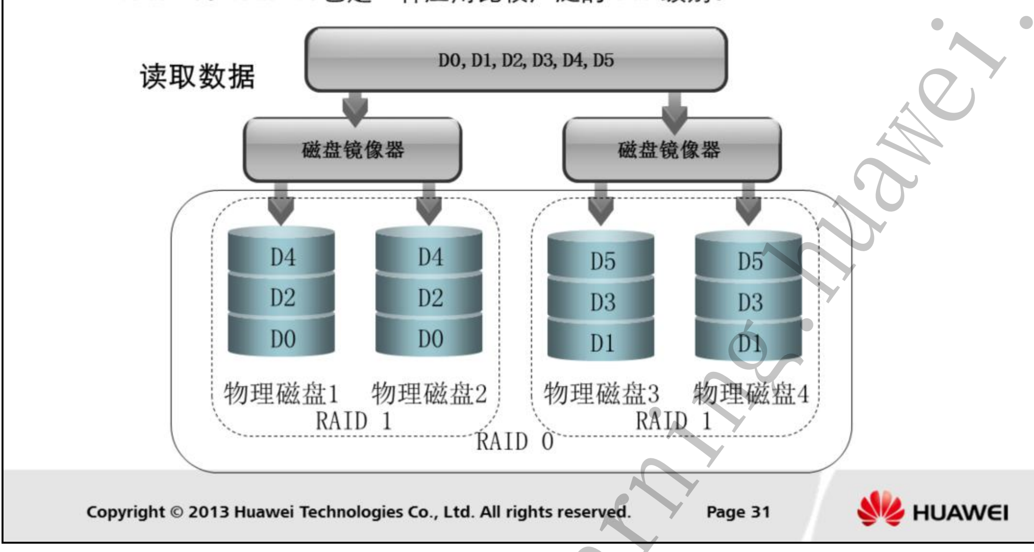
  例：DP0=D0 XOR D5 XOR D10XOR D15



 Page 30

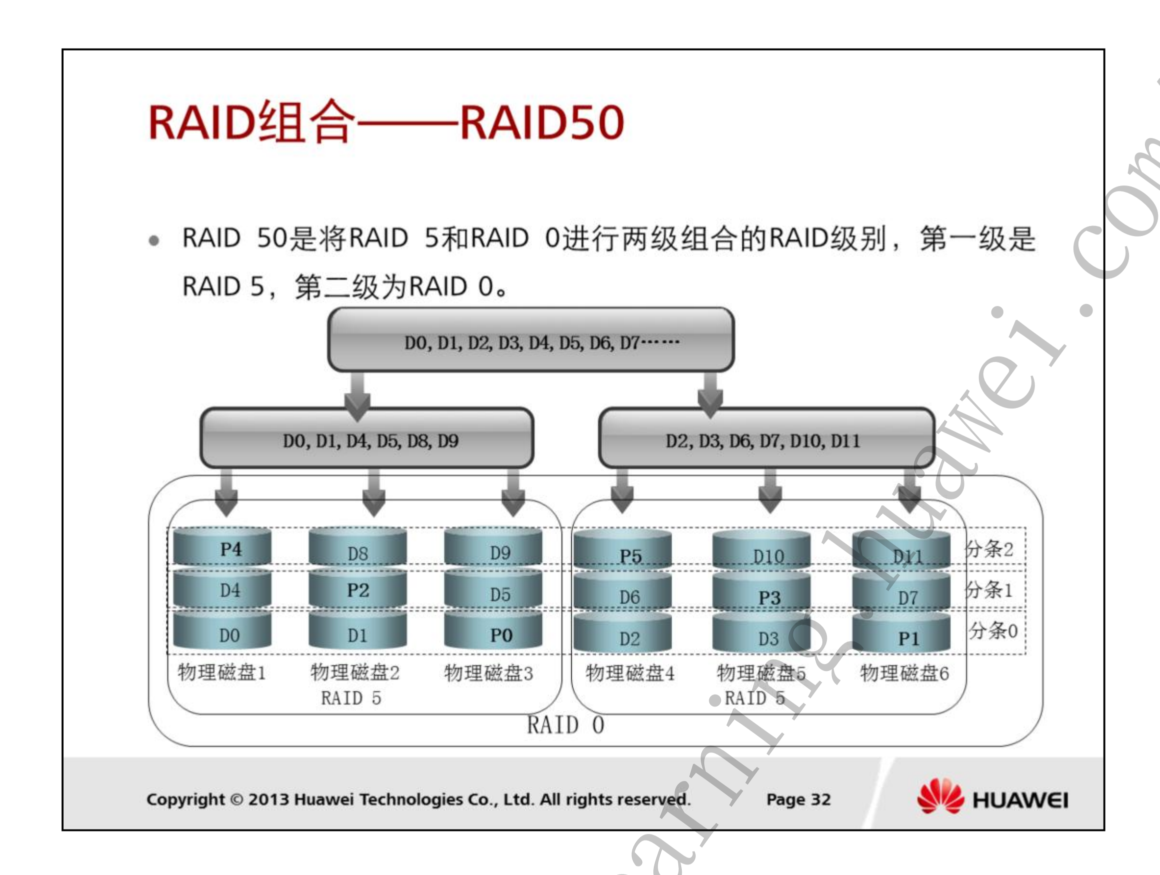


根据DP3和斜向校验恢复出D8， D8 =D3 ⊕ P1 ⊕DP3 ⊕ D13，利用P2和横向校验恢复出D9， D9 =D8 ⊕ D10 ⊕ D11 ⊕ P2；根据DP4和斜向校验恢复出D4，利用P1和横向校验恢复出D5 ……进而恢复出磁盘1、2上的所有数据。



RAID 10集RAID 0和RAID 1的优点于一身，适合应用在速度和容错要求都比较高的场合。先进行镜像，再进行分条。物理磁盘1和物理磁盘2组成RAID 1，物理磁盘3和物理磁盘4组成RAID 1，两个RAID 1再进行RAID 0。

当不同RAID1中的磁盘，如物理磁盘2和物理磁盘4发生故障导致数据失效时，整个阵列的数据读取不会受到影响，因为物理磁盘1和物理磁盘3上面已经保存了一份完整的数据。但是如果组成RAID 1的磁盘（如物理磁盘1和物理磁盘2）同时故障，数据将不能正常读取。



RAID 50是RAID 5和RAID 0的结合，先将3个或3个以上磁盘实现RAID 5，再把若干个RAID 5进行RAID 0分条。 RAID 50需要至少6个磁盘构成，把数据分条后分放到各个RAID 5中，在RAID 5中再进行分条和计算校验值的分布式存储。

物理磁盘1、2、3实现RAID 5，物理磁盘4、5、6实现RAID 5，再将两个RAID 5放在一起进行分条。允许不同RAID 5中的多块磁盘同时失效，但是一旦同一RAID 5中的两块磁盘故障，将会导致阵列失效。

常用RAID级别的比较

RAID级别	RAID 0	RAID 1	RAID 3	★RAID 5	★RAID6	★RAID 10	RAID 50
容错性	无	有	有	有	有	有	有
冗余类型	无	复制	奇偶校验	奇偶校验	奇偶校验	复制	奇偶检验
热备盘选项	无	有	有	有	有	有	有
读性能	高	低	高	高	高	一般	高
随机写性能	高	低	最低	低	低	一般	低
连续写性能	高	低	低	低	低	一般	低
最小硬盘数	2块	2块	3块	3块	4 块	4块	6块
可用容量	N * 单块硬盘容量	(1/N) * 单块硬盘容量	(N -1) * 单块硬盘容量	(N -1) * 单块硬盘容量	(N -2) * 单块硬盘容量	(N /2) * 单块硬盘容量	(N -2) * 单块硬盘容量

Page 33



RAID组成员盘个数不建议过多，例如超过20块成员盘的RAID,不但性能比8-9块成员盘RAID组(建议RAID5成员盘个数)低，且在运行过程中RAID组失效的概率增加。

RAID 典型应用场景

RAID级别	RAID 0	RAID 1	RAID 3	RAID 5 / 6	RAID 10	RAID50
典型应用环境	迅速读写，安全性要求不高，如图形工作站等	随机数据写入，安全性要求高，如服务器、数据库存储领域	连续数据传输，安全性要求高，如视频编辑、大型数据库等	随机数据传输，安全性要求高，如邮件服务器，文件服务器等	数据量大，安全性要求高，如银行、金融等领域	随机数据传输，安全性要求高，并发能力要求高，如邮件服务器，www服务器等。



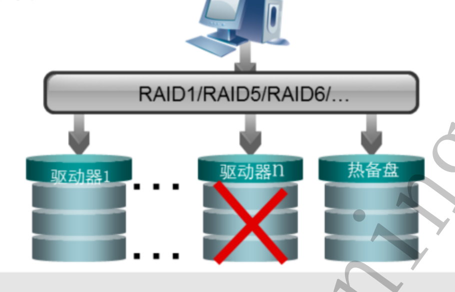


目录



热备盘

- 热备（Hot Spare）：当冗余的RAID阵列中某个磁盘失效时，在不干扰当前RAID系统正常使用的情况下，用RAID系统中另外一个正常的备用磁盘顶替失效磁盘。
- 热备通过配置热备盘实现，热备盘分为全局热备盘和局部热备盘。
- 热备盘要求和RAID组成员盘的容量，接口类型，速率一致，最好是采用同一厂家的同型号硬盘。



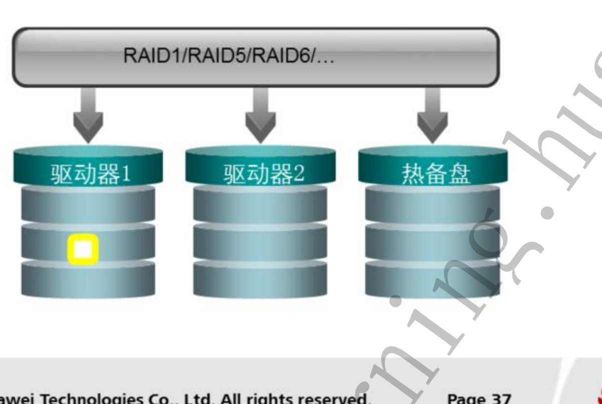
 Page 36



- 局部热备盘（Dedicated Spare）
  - 适用范围：所有冗余RAID组;
  - 特点：某RAID组专有；仅在使用时才占用实际的物理磁盘空间,其spare空间分布在指定的磁盘上;
- 全局热备盘（Global Spare）
  - 适用范围：所有冗余RAID组;
  - 特点：所有RAID组共有；仅在使用时才占用实际的物理磁盘空间；其spare空间分布在指定的磁盘上。

预拷贝

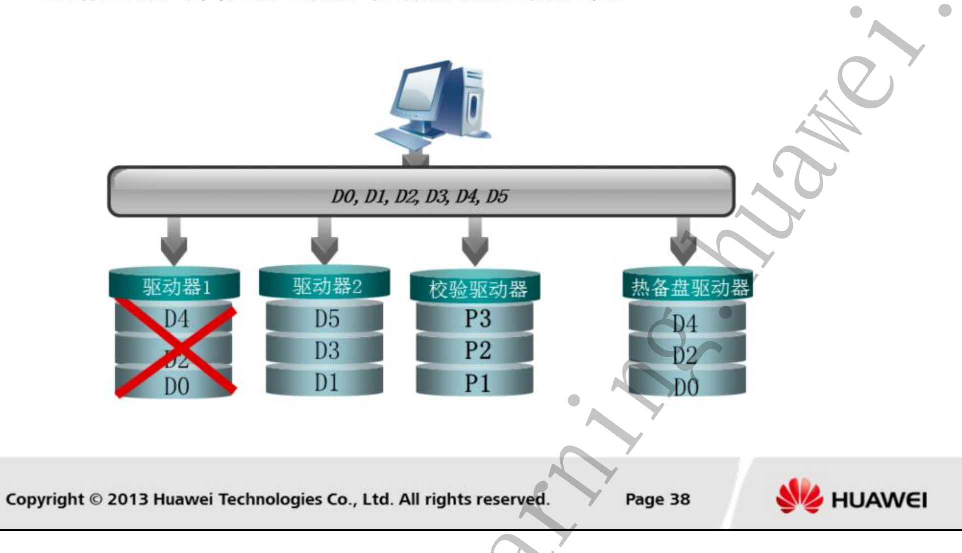
- 预拷贝：系统通过监控发现RAID组中某成员盘即将故障时，将即将故障成员盘中的数据提前拷贝到热备盘中，有效降低数据丢失风险。



 Page 37



在阵列运行的过程中，可通过磁盘健康状态检测，得出一系列参数然后综合分析该磁盘的健康状态，然后再根据这个健康状态看是否执行预拷贝的动作。

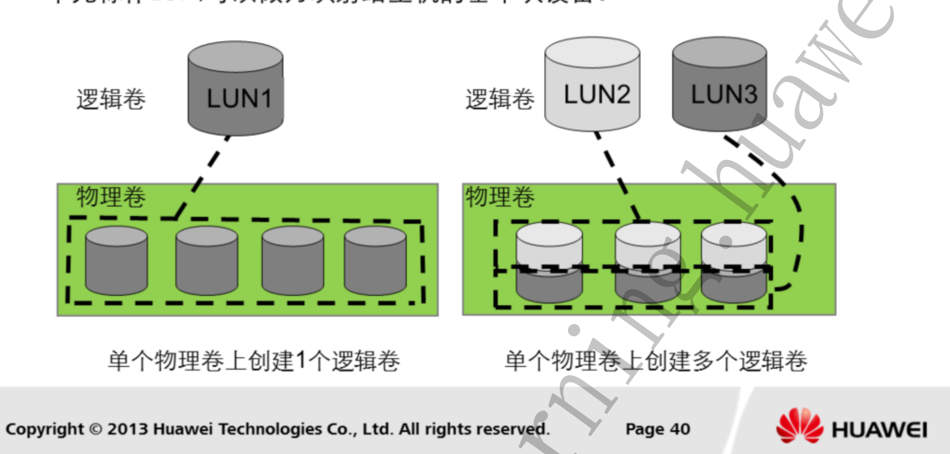


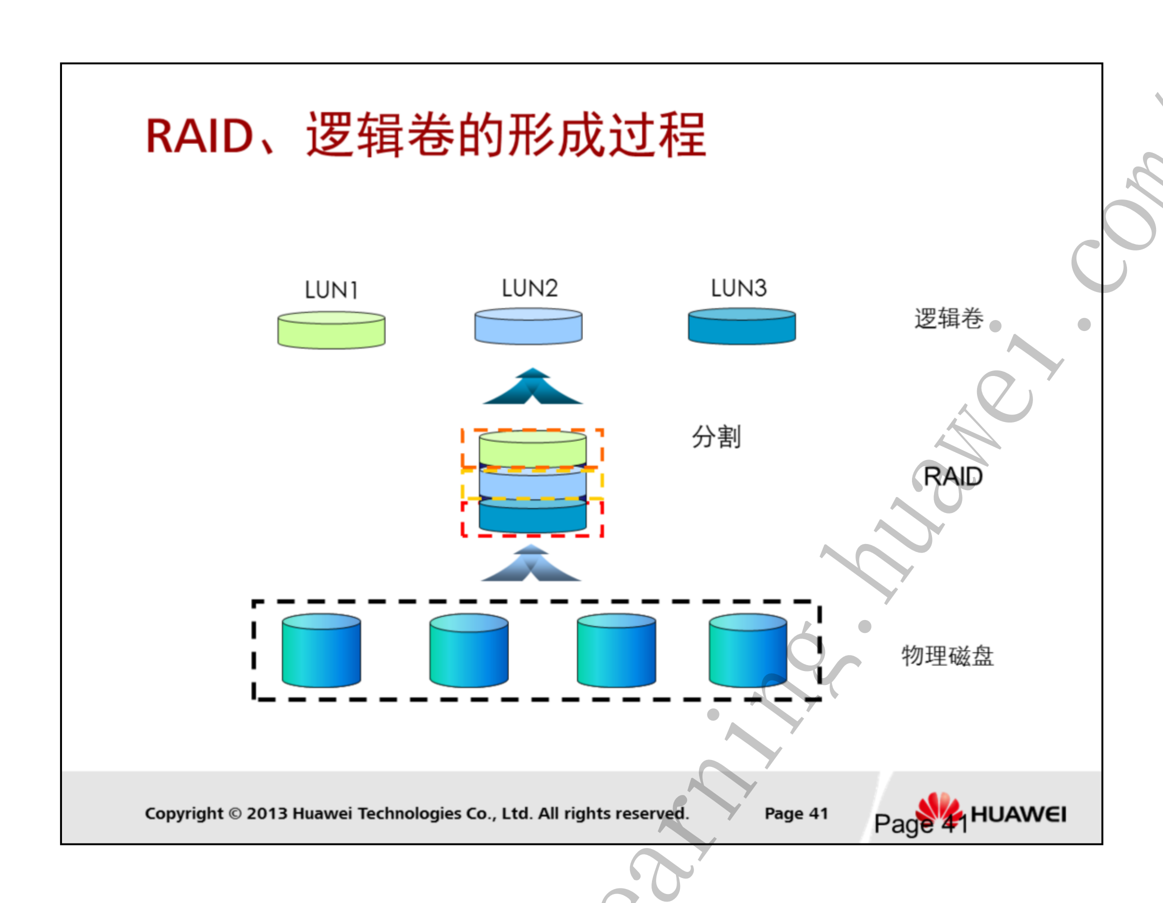
- 重构需满足以下条件：
  - 有一个热备份盘存在
  - 所有磁盘都配置为冗余阵列(RAID 1,3,5,6,10,50)
- 重构的所有操作都是在不中断系统操作的情况下进行的。



目录




RAID与LUN的关系
RAID由几个硬盘组成 ，从整体上看相当于由多个硬盘组成的一个大的物理卷。
在物理卷的基础上可以按照指定容量创建一个或多个逻辑单元，这些逻辑单元称作LUN,可以做为映射给主机的基本块设备。
逻辑卷
LUN1
物理卷
单个物理卷上创建1个逻辑卷
逻辑卷
LUN2
LUN3
物理卷
单个物理卷上创建多个逻辑卷
Copyright © 2013 Huawei Technologies Co., Ltd. All rights reserved.
Page 40
HUAWEI

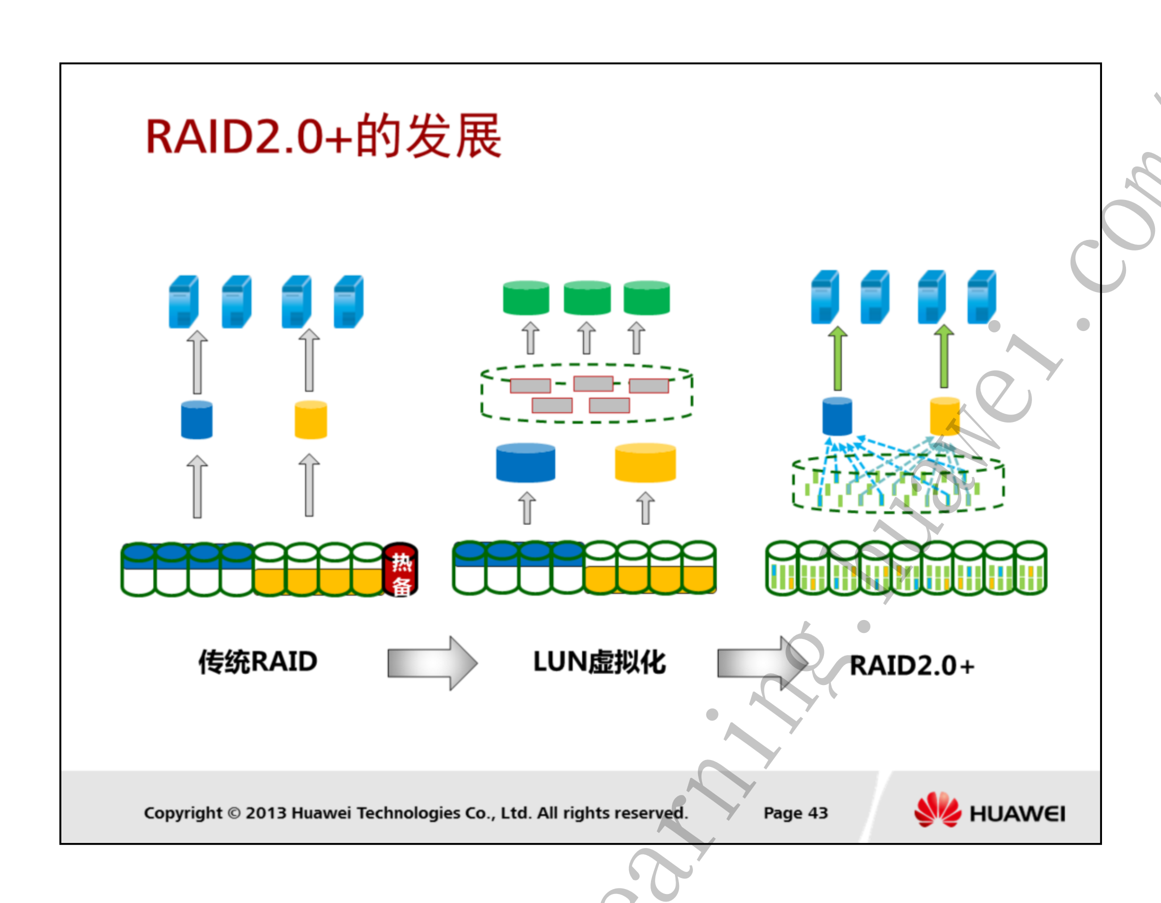


首先由多块物理磁盘创建RAID，再根据业务的需要，在RAID的基础上划分LUN,将LUN映射给服务器端，服务器端看到的是磁盘的形式，而不是LUN。

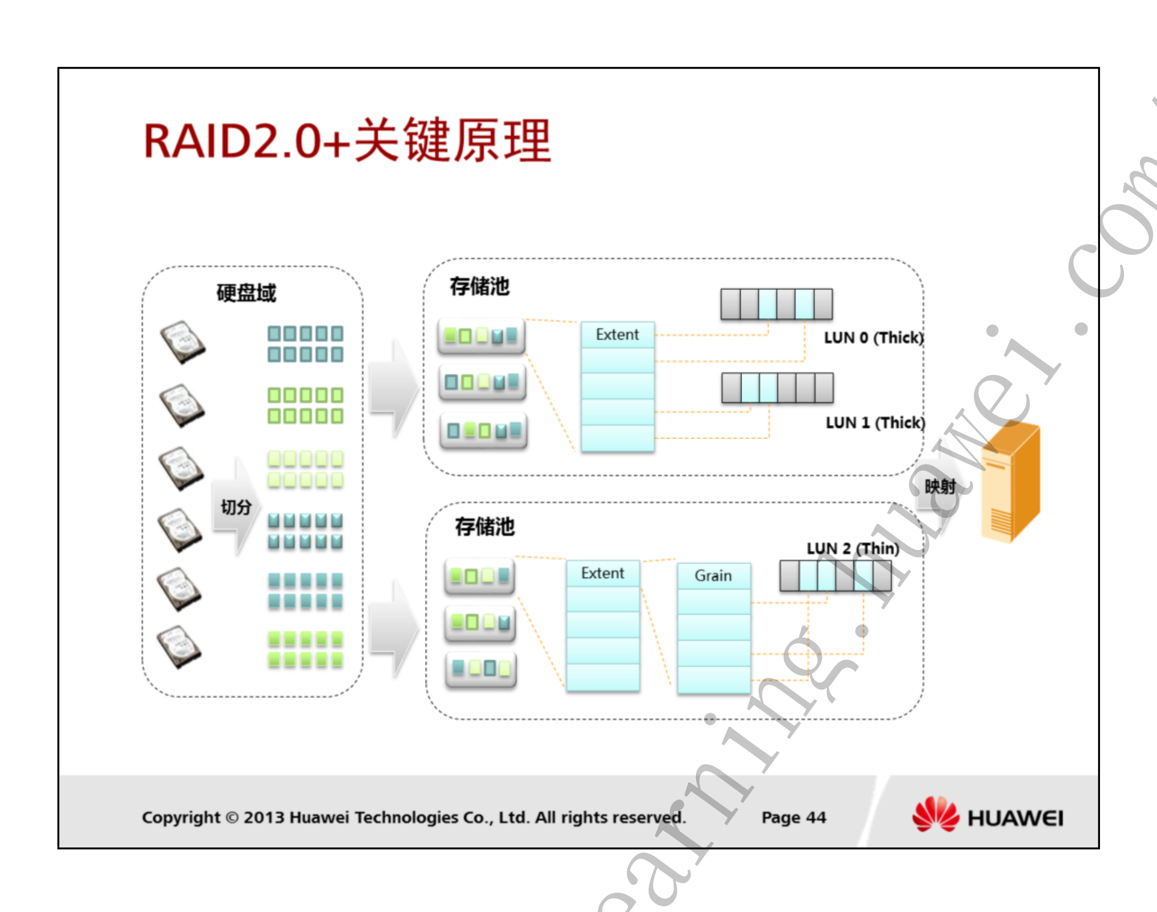


目录





在最初的RAID技术中，是将几块小容量廉价的磁盘组合成一个大的逻辑磁盘给大型机使用。后来硬盘的容量不断增大，组建RAID的初衷不再是构建一个大容量的磁盘，而是利用RAID技术实现数据的可靠性和安全性，以及提升存储性能，由于单个容量硬盘都已经较大了，数据硬盘组建的RAID容量更大，然后再把RAID划分成一个一个的LUN映射给服务器使用。随着硬盘技术的发展，单块硬盘的容量已经达到数T，传统RAID技术在重建的过程中需要的时间越来越长，也增加了在重构过程中其它硬盘再坏掉对数据丢失造成的风险，为了解决这一问题，块虚拟化技术应运而生，将以前以单块硬盘为成员盘的RAID技术再细化，将硬盘划分成若干的小块，再以这些小块为成员盘的方式构建RAID,也就是现在业界所说的RAID2.0+技术。



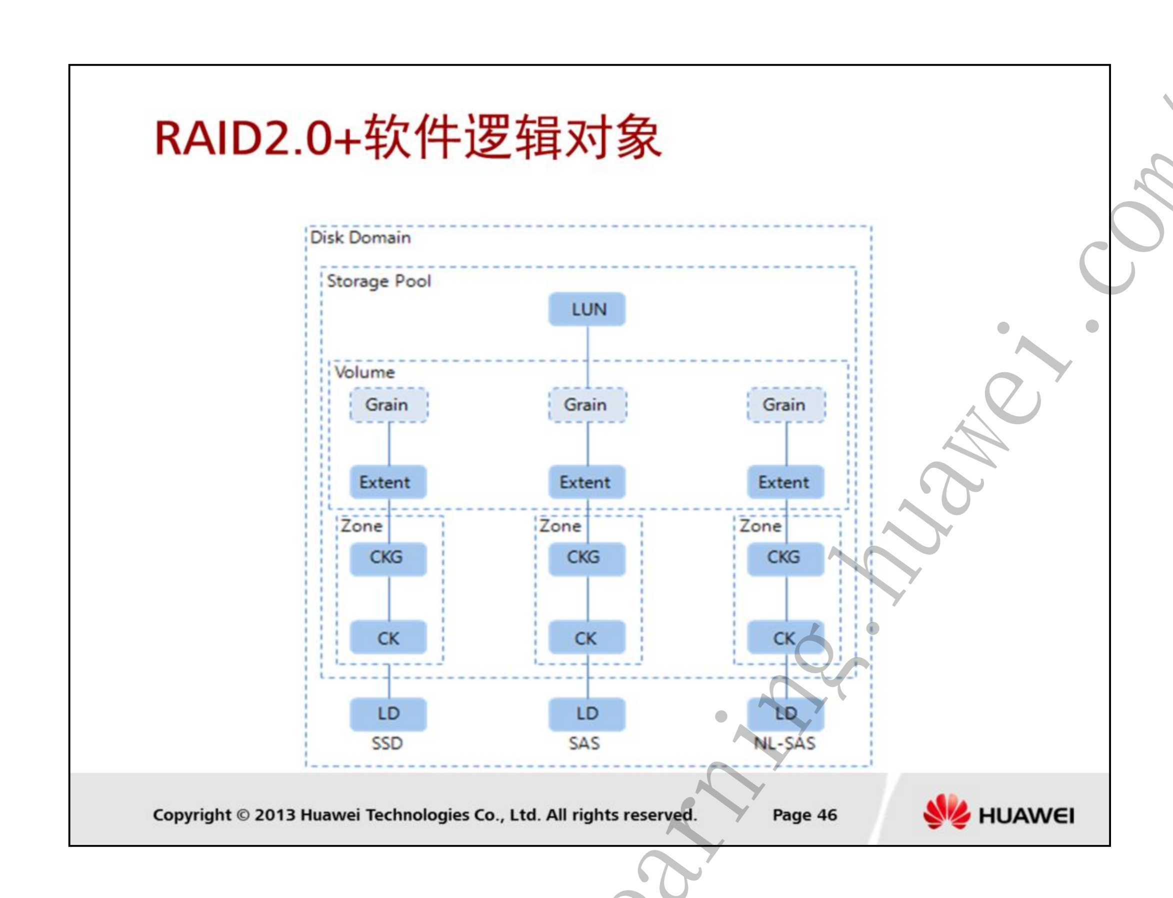
在RADI2.0+中，把硬盘域内每个硬盘切分为固定64MB的块（CK）,硬盘域内同种类型的硬盘被划分为一个个的Disk Group（DG），从同一个DG上随机选择多个硬盘，每个硬盘选取CK按照RAID算法组成Chunk Group（CKG）,CKG被划分为固定大小的Extent, Thick LUN以Extent为单位映射到LUN, Grain在Extent的基础上进行更细粒度的划分,Thin LUN以Grain 为单位映射到LUN.

RAID2.0+软件逻辑对象

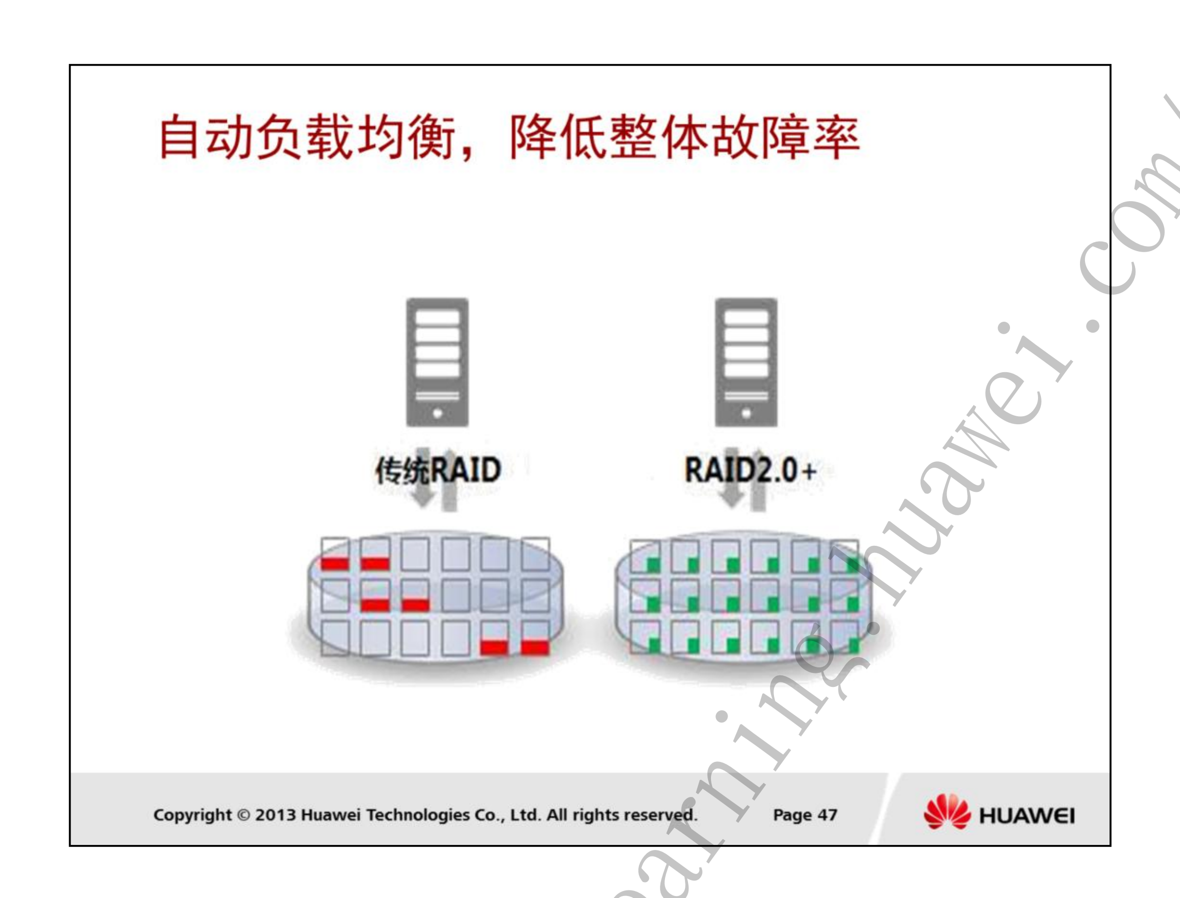
Disk Domain
Storage Pool
LUN
Volume
Grain　Grain　Grain
Extent　Extent　Extent
Zone　Zone　Zone
CKG　CKG　CKG
CK　CK　CK
LD　LD　LD
SSD　SAS　NL-SAS

 Page 45 HUAWEI

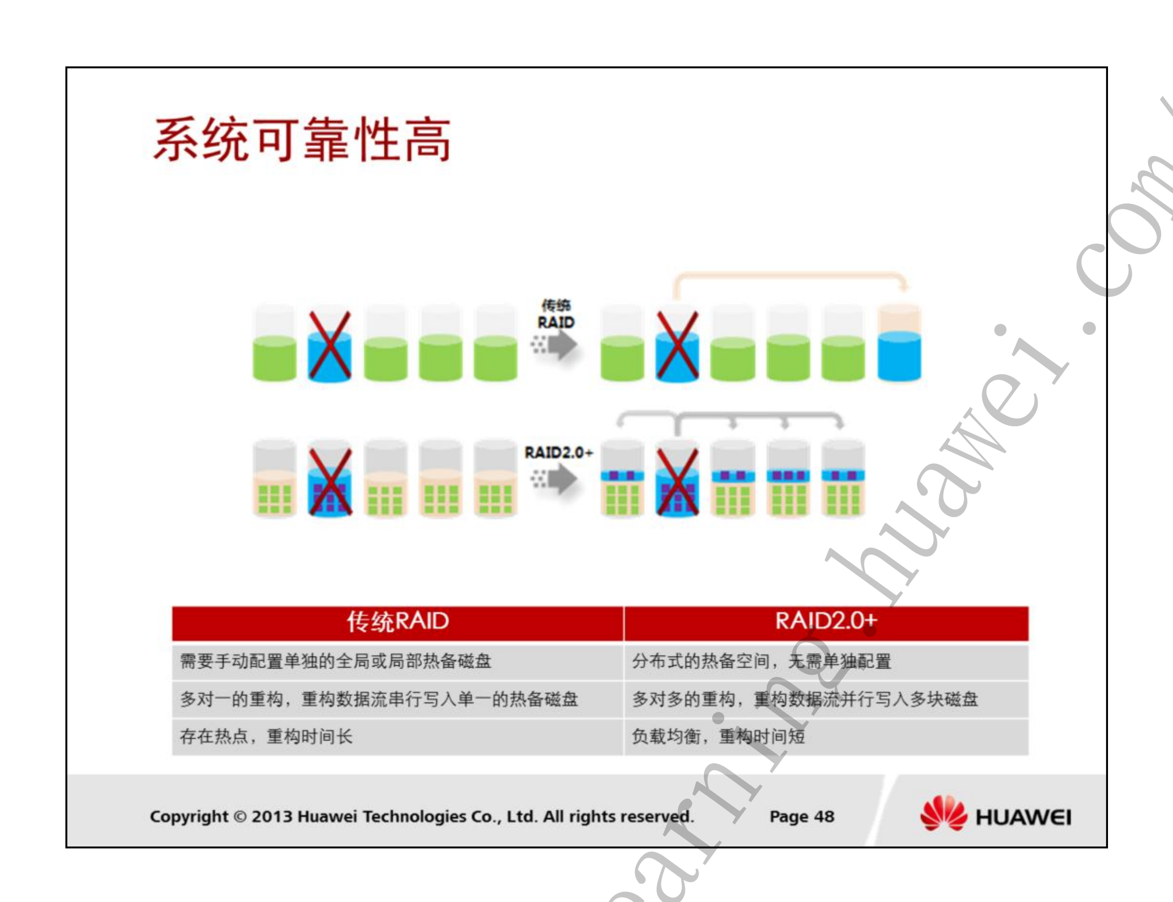
- Disk Domain（磁盘域），一个硬盘域上可以创建多个存储池（Storage Pool）一个硬盘域的硬盘可以选择SSD、SAS、NL-SAS中的一种或者多种，不同硬盘域之间是完全隔离的，包括故障域、性能和存储资源等。
  - Storage Pool（存储池）& Tier，一个存储池基于指定的一个硬盘域创建，可以从该硬盘域上动态的分配Chunk（CK）资源，并按照每个存储层级（Tier）的“RAID策略”组成Chunk Group（CKG）向应用提供具有RAID保护的存储资源
  - Disk Group（DG），由硬盘域内相同类型的多个硬盘组成的集合，硬盘类型包括SSD、SAS和NL-SAS三种。
  - LD（逻辑磁盘），是被存储系统所管理的硬盘，和物理硬盘一一对应。
  - Chunk（CK），是存储池内的硬盘空间切分成若干固定大小的物理空间，每块物理空间的大小为64MB，是组成RAID的基本单位。
  - Chunk Group（CKG），是由来自于同一个DG内不同硬盘的CK按照RAID算法组成的逻辑存储单元，是存储池从硬盘域上分配资源的最小单位。
  - Extent，是在CKG基础上划分的固定大小的逻辑存储空间，大小可调，是热点数据统计和迁移的最小单元（数据迁移粒度），也是存储池中申请空间、释放空间的最小单位。
  - Grain，在Thin LUN模式下，Extent按照固定大小被进一步划分为更细粒度的块，这些块称之为Grain,Thin LUN以Grain为粒度进行空间分配，Grain内的LBA是连续的。Thin LUN以Grain为单位映射到LUN，对于Thick LUN，没有该对象。



- Volume & LUN，Volume即卷，是存储系统内部管理对象；LUN是可以直接映射给主机读写的存储单元，是Volume对象的对外体现。



- 数据在存储池中硬盘上的自动均衡分布，避免了硬盘的冷热不均，从而降低了存储系统整体的故障率。



- 华为OceanStor 18000系列存储系统针对硬盘采用了多重故障容错设计，具有硬盘在线诊断、DHA（Disk Health Analyzer，硬盘故障诊断与预警）、坏道后台扫描、坏道修复等多种可靠性保障，RAID2.0+技术会根据热备策略自动在存储池中预留一定数量的热备空间，用户无需进行设置，当系统自动检测到硬盘上某个区域不可修复的介质错误或整个硬盘发生故障时，系统会自动进行重构，将受影响的数据块数据快速重构到其他硬盘的热备空间中，实现系统的快速自愈合。

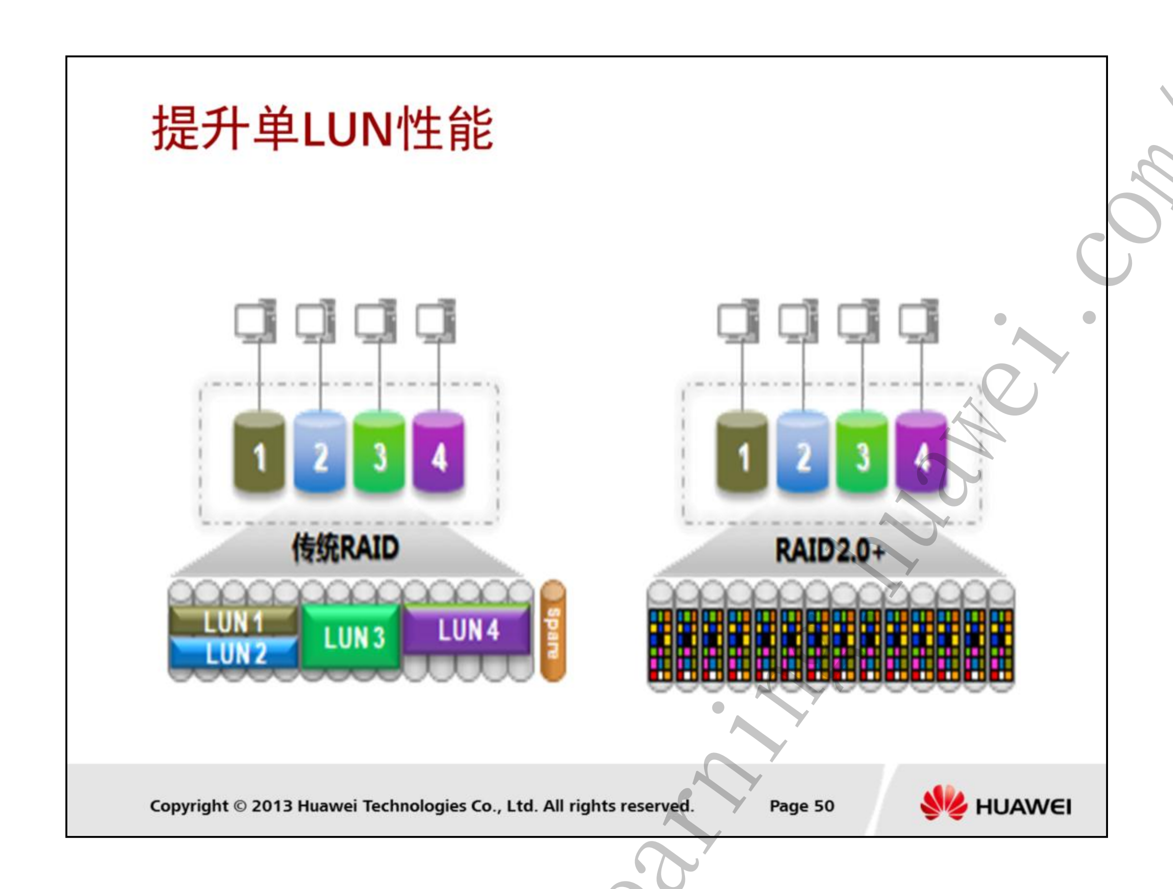
快速精简重构，改善双盘失效率

传统RAID重构原理（多对一）

RAID2.0+重构原理（多对多）

Page 49 HUAWEI

- 左图传统RAID中，HDD0～HDD4五块硬盘创建RAID5，HDD5为热备盘，当HDD1故障后，HDD0、HDD2、HDD3、HDD4通过异或算法将重构的数据写入HDD5中；
- 在右图的RAID2.0+示意图中，当HDD1故障后，故障盘HDD1中的数据按照CK的粒度进行重构，只重构已分配使用的CK（图中HDD1的 和 ），存储池中所有的硬盘都参与重构过程，重构的数据分布在多块硬盘中（图中的HDD4和HDD9）。
- 重构速率的提升还得益于RAID2.0+技术对故障的处理更加精细有效，RAID2.0+在原有坏道修复和全盘失效重构两级故障修复之间增加了数据块的故障修复，能够基于块（CK）的粒度只重构已分配并使用了的空间，通过对实际使用空间的有效识别，当硬盘出现故障时，RAID2.0+能够通过精简重构进一步缩短重构时间，降低数据丢失的风险。



- 传统存储的RAID通常是以单个磁盘为粒度来建立RAID，RAID被限制在有限的几个磁盘上，不能充分发挥每个磁盘的所有资源。所以当主机对一个较小的卷进行密集访问时，只能访问到有限的几个磁盘，这就造成磁盘访问瓶颈，导致磁盘热点。
- 而RAID2.0+技术基于Chunk而非物理磁盘构成RAID。一个物理磁盘上的不同CK可以用于构成不同RAID类型的卷。这种基于条块（CK）的实现，可以在同一个物理磁盘上实现不同的RAID类型，为每个卷的RAID类型提供更优化的选择。
- 对于HVS阵列而言，即使是很小的卷也可以通过CK的方式分布到很多磁盘上。宽条带化技术使得小的卷不再需要额外的大容量即可获得足够的高性能，且避免了磁盘热点。物理磁盘上剩余的CK还可以用于其它的卷。



RAID2.0+基于业界领先的块虚拟化技术实现，卷上的数据和业务负荷会自动均匀分布到存储池所有的物理硬盘上，借助于智能的Smart系列效率提升套件，T系列存储系统能自动根据业务所需的性能、容量、冷热数据等因素在后台进行智能调配，灵活地适应企业业务的快速变化。



总结

- RAID分类及原理
- 各种RAID特点
- RAID在数据保护方面所采用的技术
- RAID2.0+技术





1. 您能简要描述RAID中的数据是如何组织的？分条深度对RAID的性能有影响吗?
2. RAID组的状态有哪些？它们之间是如何转换的?
3. 您能描述一下RAID5的数据组织方式和重构过程是如何实现的?
4. RAID5和RAID1的应用场景有区别吗？有哪些?
5. 在客户更关注可靠性和性能的情况下，给客户推荐合适的RAID方案有哪些?
6. RAID与LUN的关系是什么?

 Page 53



练习题

多选题

1、以下阵列类型中，具有冗余功能的有（ ）

A. RAID0

B. RAID1

C. RAID3

D. RAID5

判断题

1、在RAID10中，任意两块磁盘出故障都不影响读取数据。(T or F)

2、RAID2.0+技术将硬盘划分成若干的小块，再以这些小块为成员盘的方式构建RAID。（T or F）

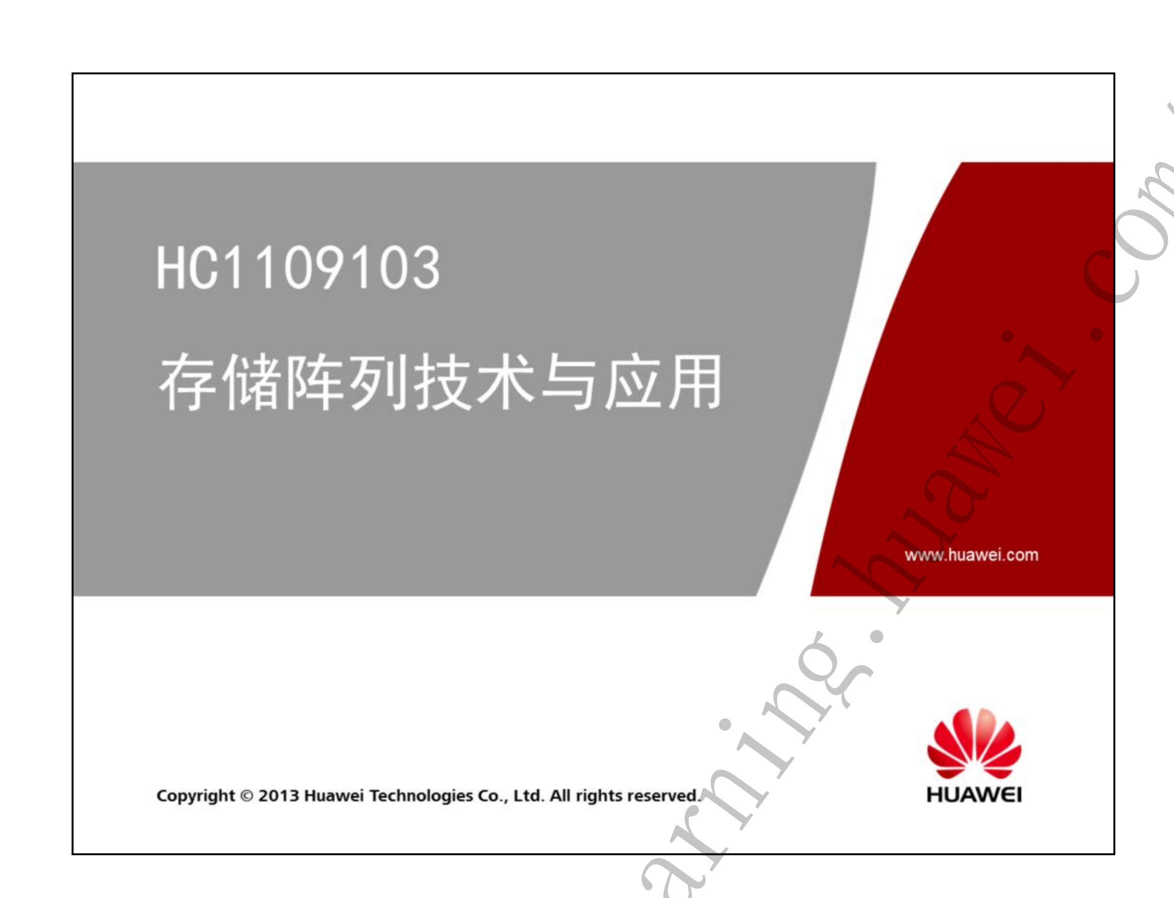
 Page 54 

- 习题答案
  - 多选题
  - 1、BCD
  - 单选题
  - 1、F
  - 2、T


Thank you
www.huawei.com
Copyright © 2013 Huawei Technologies Co., Ltd. All rights reserved.
Page 55
HUAWEI

HC1109103 存储阵列技术及应用




HC1109103
存储阵列技术与应用
www.huawei.com
Copyright © 2013 Huawei Technologies Co., Ltd. All rights reserved.
HUAWEI



目标

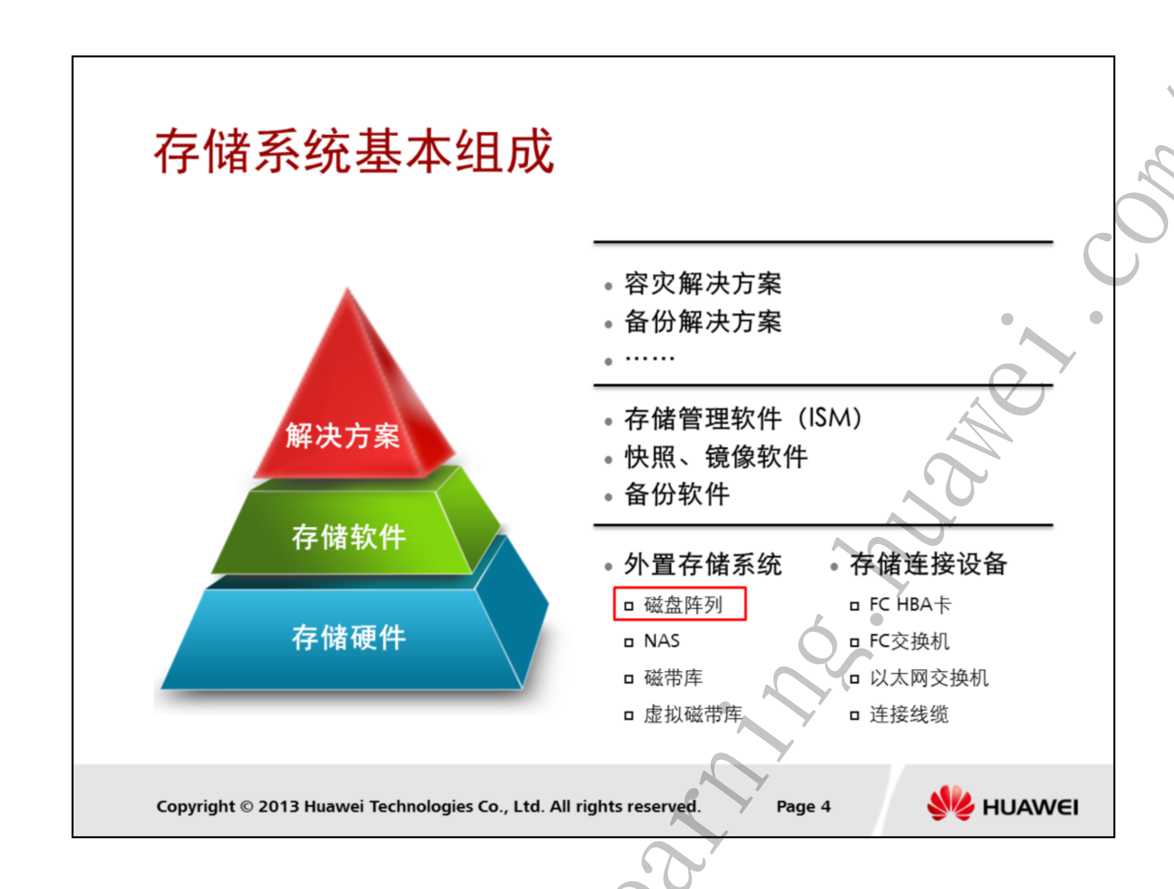
- 学习完本章节后，您将能够：
  - 熟悉存储阵列系统的组成
  - 掌握华为存储阵列的技术
  - 掌握华为存储阵列基础配置



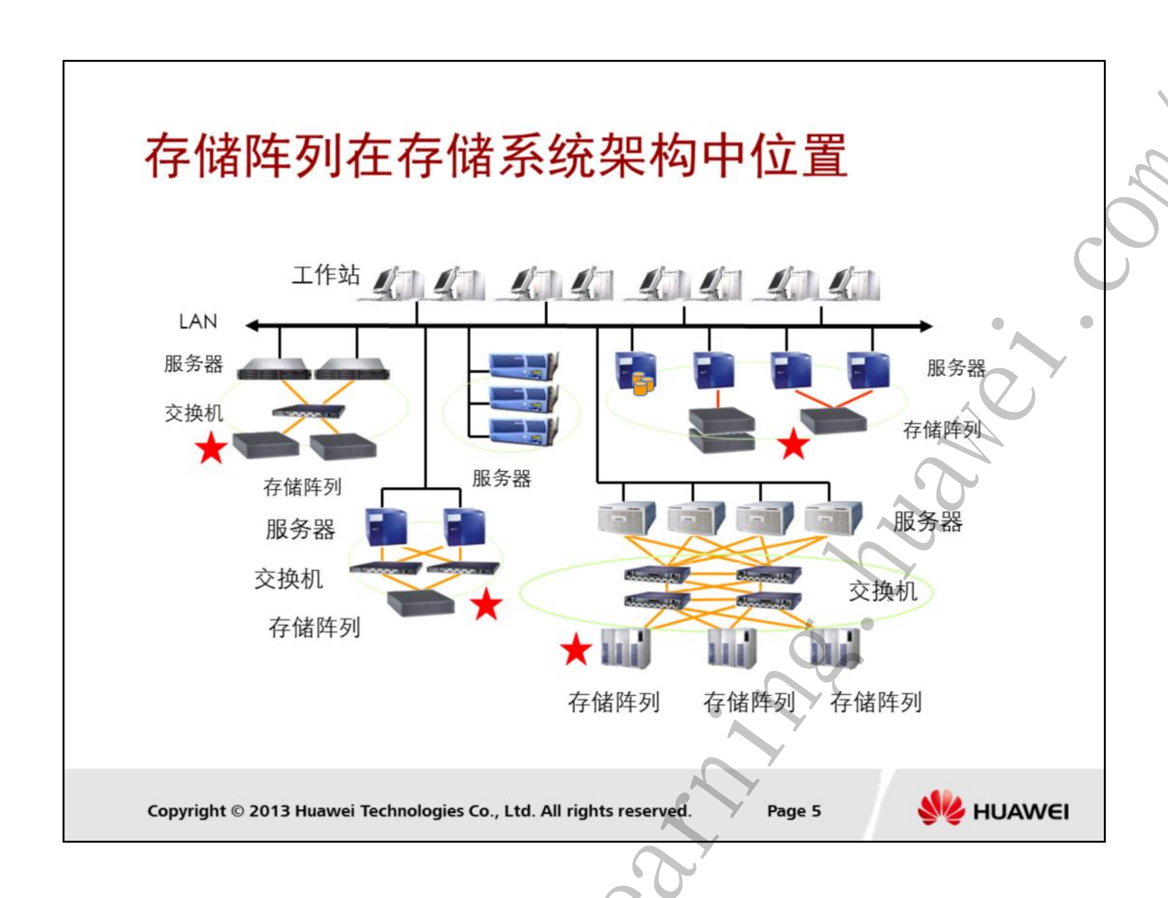


目录





当今的存储技术不是一个单独而孤立的技术，实际上，完整的存储系统应该是由一系列组件构成。目前，人们把存储系统分为了硬件架构部分、软件组件部分以及实际应用时的存储解决方案部分。硬件部分又分为外置的存储系统，主要是指人们实际的存储设备，比如磁盘阵列、磁带库等。因为软件组件的存在，使存储设备的可用性得到了大大的提高，数据的镜像、复制，自动的数据备份等等数据操作都可以通过对存储软件的控制来完成。一个设计良好的存储解决方案，是使人们数据存储工作更加简单易行的最佳保障，设计优秀的存储解决方案，不仅可以使存储系统实际部署的时候更简单容易，更可以降低客户的总体拥有成本（TCO），使客户的投资能得到良好的保护。



在存储系统架构中，磁盘阵列做为数据存储设备，为用户业务系统提供数据存储服务。存储阵列设备成为用户数据业务稳定、可靠、高效运行的重要因素。

存储阵列系统硬件组成

- 控制框
  - 控制框用于处理各种存储业务，并管理级联在控制框下面的硬盘框。
- 硬盘框
  - 硬盘框主要用于容纳各种硬盘，为应用服务器提供充足的存储空间。

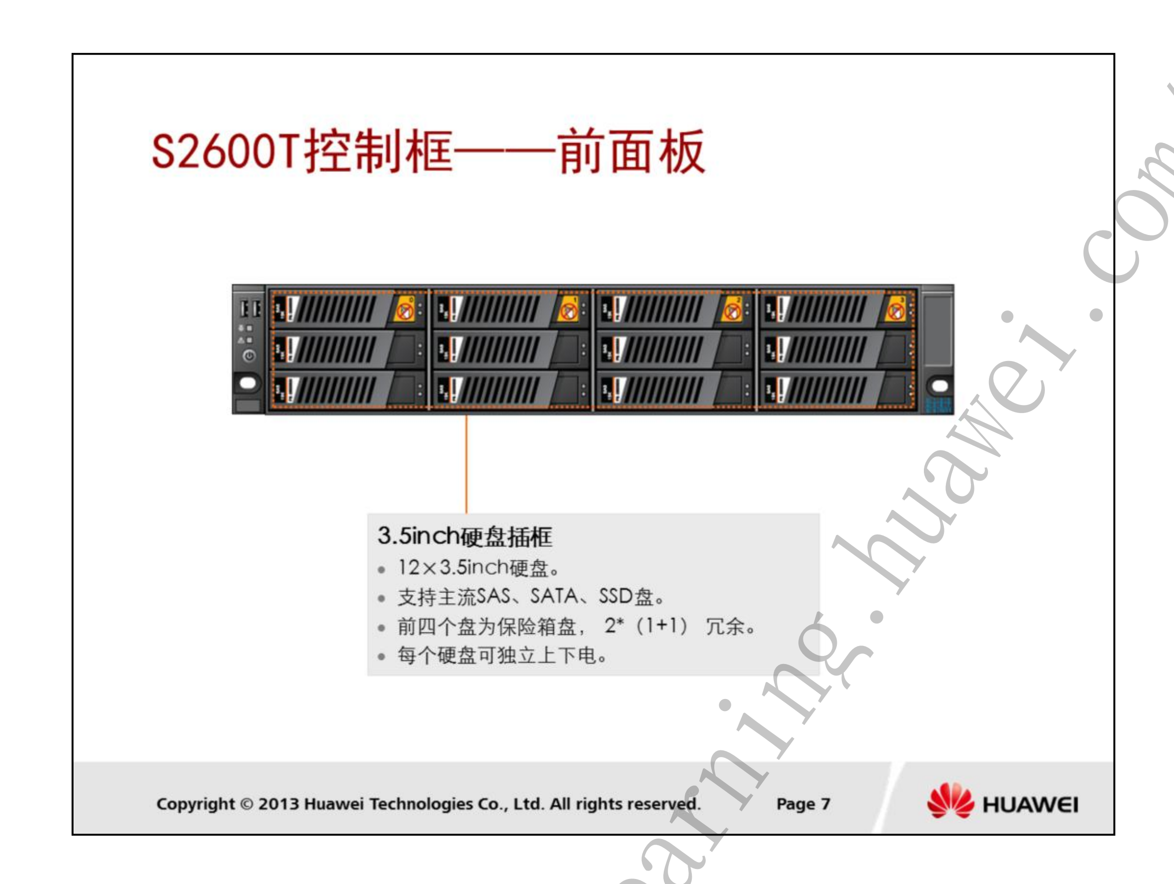




 Page 6 

存储阵列系统主要由控制框和硬盘框两部分组成，为客户提供一个高可靠、高性能、大容量的智能化存储系统。

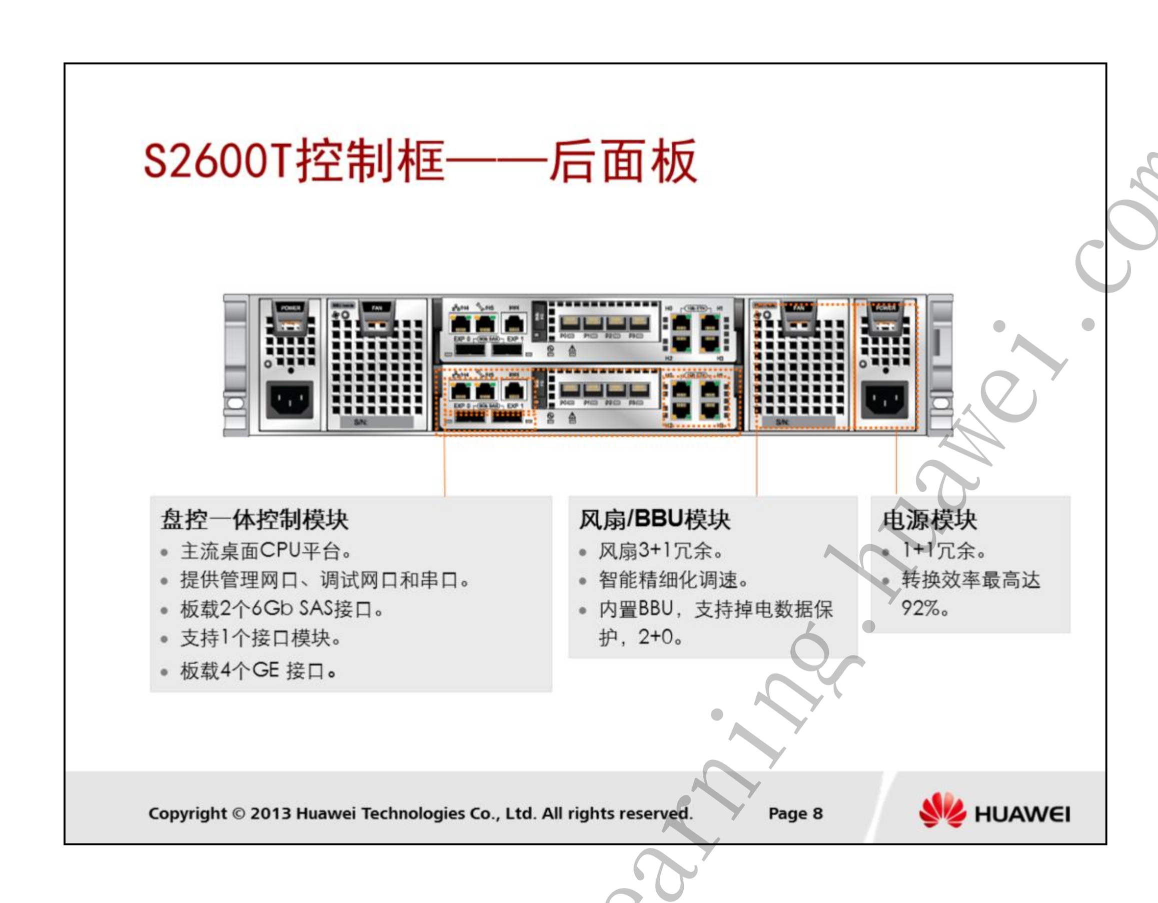
华为T系列存储阵列产品包括：S2200T/S2600T/S5500T/S5600T/S5800T/S6800T，其中S2200T/S2600T/S5500T为2U盘控一体化架构，S5600T/S5800T/S6800T为4U盘控分离架构。本课程内容将以S2600T/S5500T/S5600T为例，重点介绍盘控一体和盘控分离两种硬件形态的硬件组成和软件架构。



S2600T控制框采用部件模块化设计，主要由系统插框、控制器模块、风扇-BBU模块、电源模块和硬盘模块组成。控制框可使用交流电源模块或直流电源模块进行供电。

12盘位控制框硬盘槽位从左到右，再从上到下，按0~11进行编号，保险箱盘位于0~3号槽位。

- 保险箱盘：对于盘控一体的控制框，存储系统中控制框的前4个硬盘规划为保险箱盘；
- 保险箱盘用于存放系统重要数据，以及在电源模块故障时保存Cache中的数据；
- 每一块保险箱盘上用于存放系统重要数据的容量为23 GB，4块保险箱盘共占用92 GB的容量；
- A控2块保险箱盘，B控两块保险箱盘，分别为1+1备份。

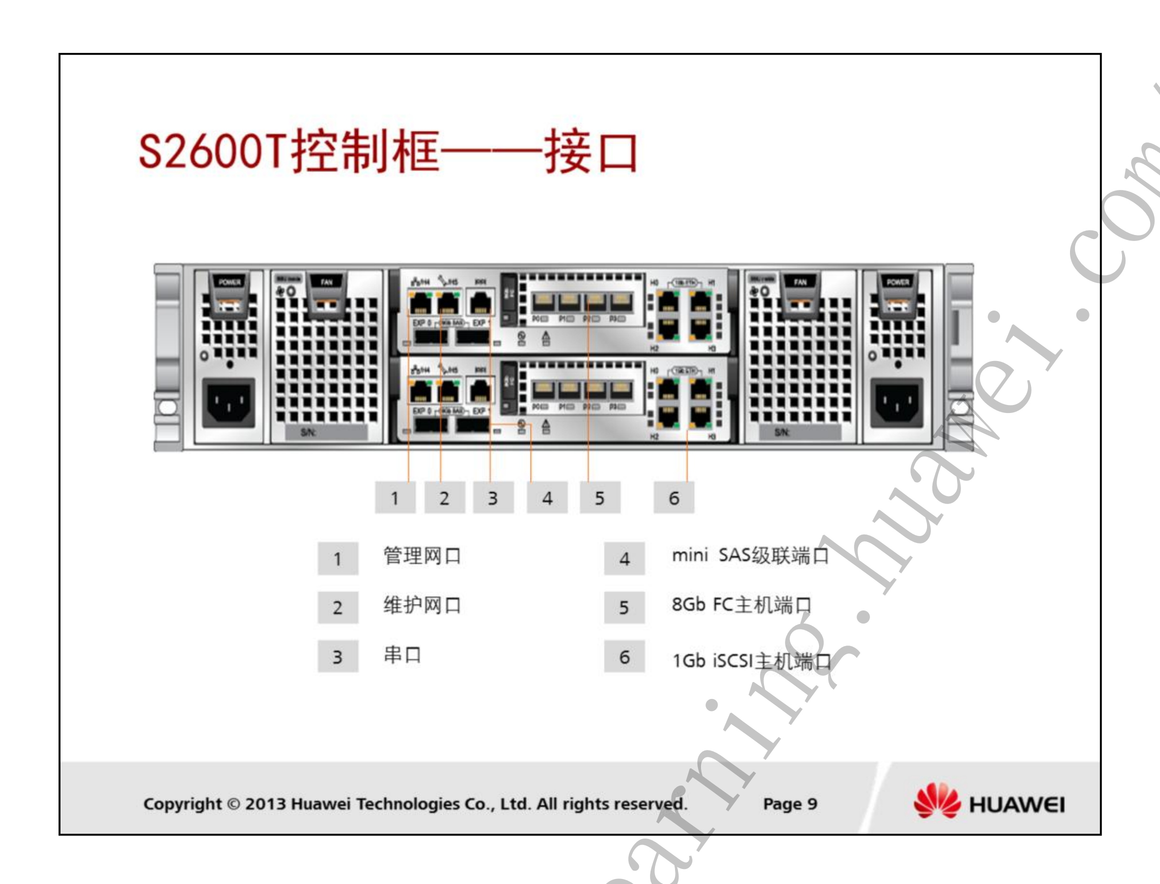


控制器模块是控制框中的核心部件，主要负责处理存储业务、接收用户的配置管理命令并保存配置信息、接入硬盘管理。

BBU模块能够在系统外部供电失效的情况下，提供后备电源支持，以保证存储系统中业务数据的安全性。在系统电源输出正常时BBU处于备份状态，当外部电源断开时，BBU能够继续给存储系统供电一段时间，使存储系统能够将Cache中的数据同步到保险箱盘中，避免业务数据的丢失。BBU支持失效隔离，当BBU发生故障时不会影响系统的正常运行。

设备操作注意事项：

- 系统正在走下电流程不允许插拔控制盒或者BBU；
- 10s内不可对同一槽位更换硬盘；
- 10s内不可互换SAS级联线；
- 升级一个BBU时不能插拔另一个BBU；
- 单电源供电，在插入另一个电源5秒之内不允许拔出另一个正在供电的电源或者该电源的电源线。



- 控制器通常包括以下接口：
  - 管理和维护网口
    - 管理模块为存储系统提供管理接口，主要包括管理网口、维护网口和串口。
    - 管理模块将系统配置数据、告警信息以及日志信息保存到保险箱硬盘上。
  - 主机I/O端口
    - I/O接口模块主要用于存储设备与应用服务器之间的连接
    - 存储设备支持的I/O接口模块包括FC接口模块、TOE接口模块、iSCSI接口模块及FCoE接口模块。
  - 硬盘框级联端口
    - 级联模块通过级联端口连接控制框和硬盘框，是控制框和硬盘框之间进行数据传输的连接点。

目前存储系统通常使用双控制器，当一个控制器出现故障时，存储系统仍能够为业务系统提供数据存储服务。

S2600T控制框支持1Gb iSCSI、10Gb TOE和8Gb FC主机接口模块，S2200T只支持8Gb FC主机接口模块。

S5500T控制框——前面板



2.5inch硬盘插框

- 24×2.5inch硬盘。
- 支持主流SAS、NL-SAS、SSD盘。
- 前四个盘为保险箱盘，2*（1+1）冗余。
- 每个硬盘可独立上下电。



3.5inch硬盘插框

- 12×3.5inch硬盘。
- 支持主流SAS、NL-SAS、SSD盘。
- 前四个盘为保险箱盘，2*（1+1）冗余。
- 每个硬盘可独立上下电。

 Page 10

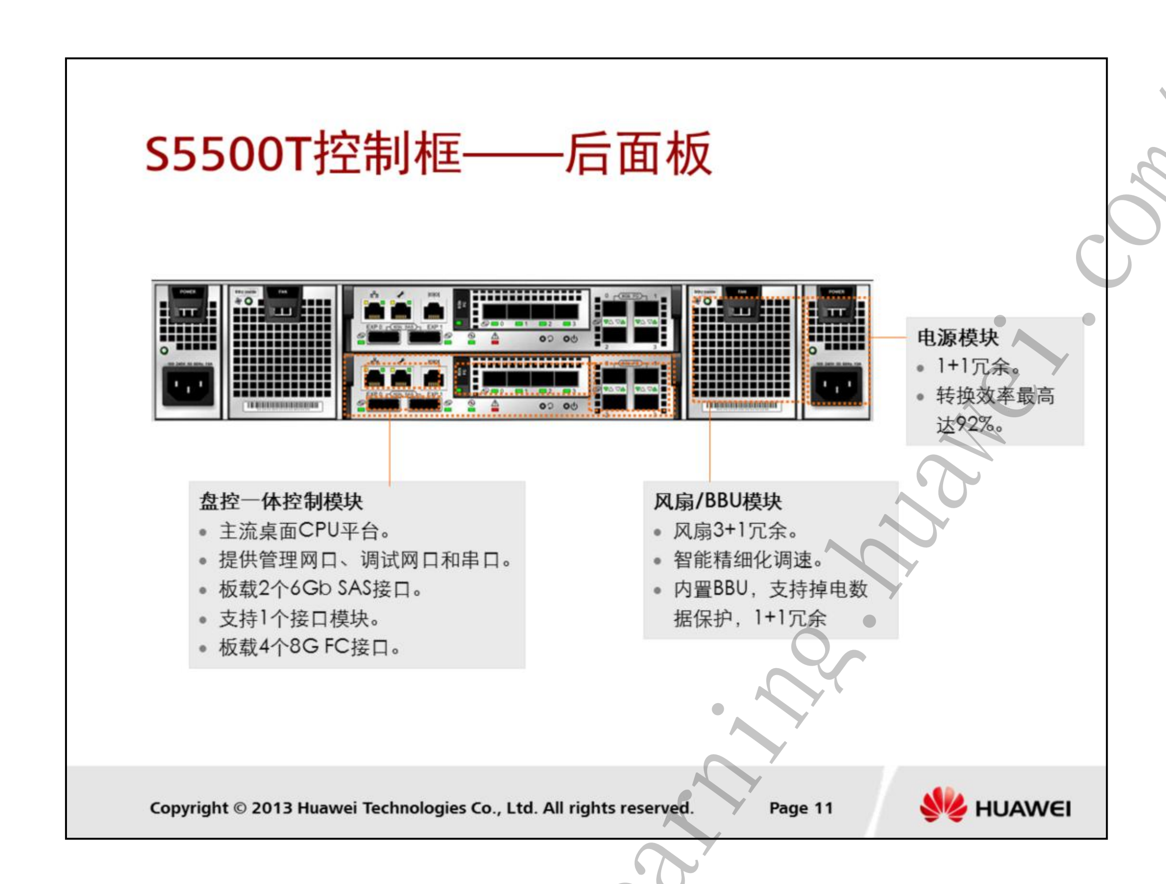


S5500T存储系统主要由控制框、硬盘框和文件引擎组成。通过控制框、硬盘框和文件引擎互联，可为客户提供一个高可靠、高性能、大容量的智能化存储系统，使存储系统实现性能监控、故障告警等功能。

控制框采用部件模块化设计，主要由系统插框、控制器模块、风扇-BBU模块、电源模块和硬盘模块组成。

控制框为2U，分为 24盘位控制框和12盘位控制框。 24盘位控制框硬盘槽位从左到右按0~23进行编号，保险箱盘位于0~3号槽位， 12盘位控制框硬盘槽位从左到右，再从上到下，按0~11进行编号，保险箱盘位于0~3号槽位。

控制框可使用交流电源模块或直流电源模块进行供电。



S5500T控制框支持1Gb iSCSI、10Gb TOE、10Gb FCoE和8Gb FC接口模块。

S5500T控制框——接口

1 2 3 4 5 6

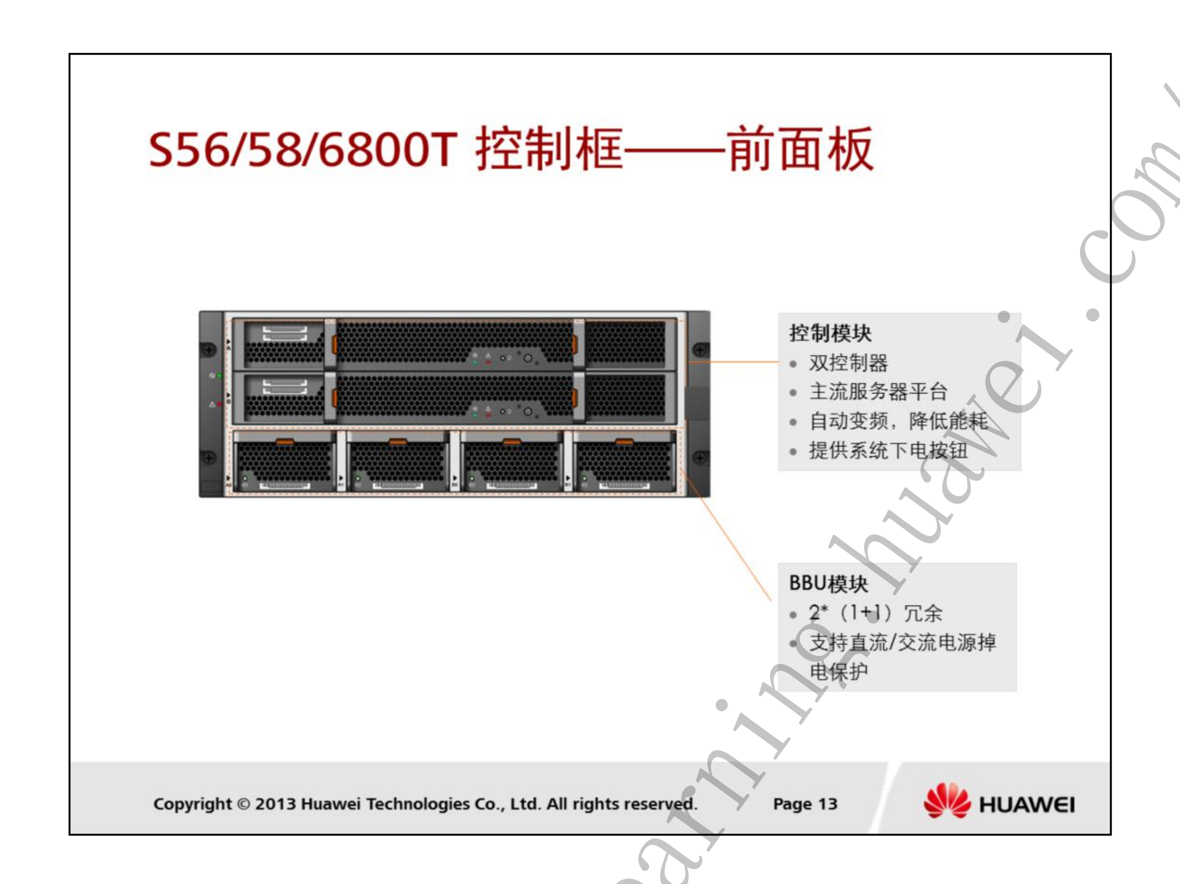
1	管理网口	4	mini SAS级联端口
2	维护网口	5	8Gb FC主机端口
3	串口	6	8Gb FC主机端口

 Page 12 HUAWEI

管理模块为存储系统提供管理接口，主要包括管理网口、维护网口和串口。管理模块将系统配置数据、告警信息以及日志信息保存到保险箱硬盘上。

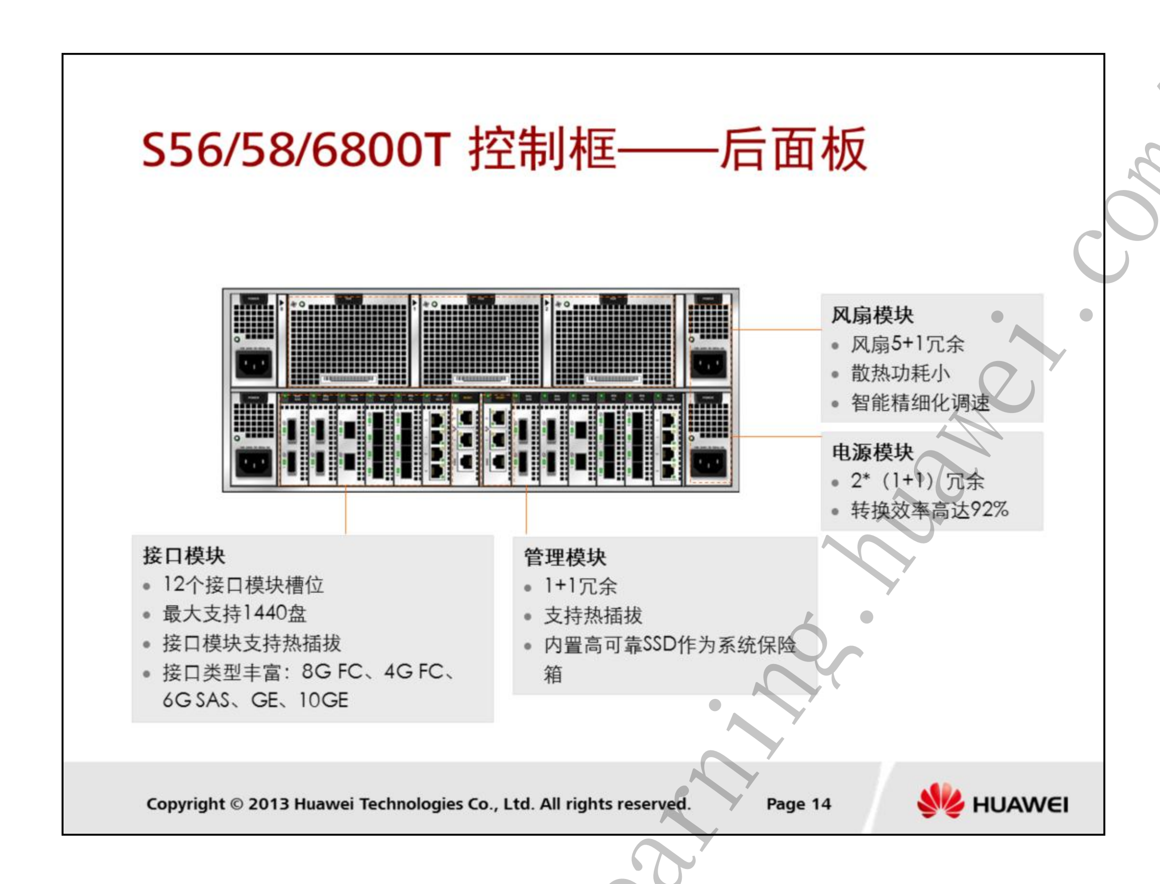
I/O接口模块主要用于存储设备与应用服务器之间的连接，目前华为存储设备支持的I/O接口模块包括8Gb FC接口模块、10Gb TOE接口模块、1Gb iSCSI接口模块以及10Gb FCoE接口模块。

级联模块通过级联端口连接控制框和硬盘框，是控制框和硬盘框之间进行数据传输的连接点。 SAS级联模块提供2个传输速率为6Gbit/s的mini SAS级联端口，用于级联硬盘框。控制框的SAS级联模块通过mini SAS电缆与存储系统的后端硬盘阵列连接。当连接的设备传输速率低于级联端口速率时，级联端口将自动适应传输速率，以保证数据传输通道的连通性和数据传输速率的一致性。

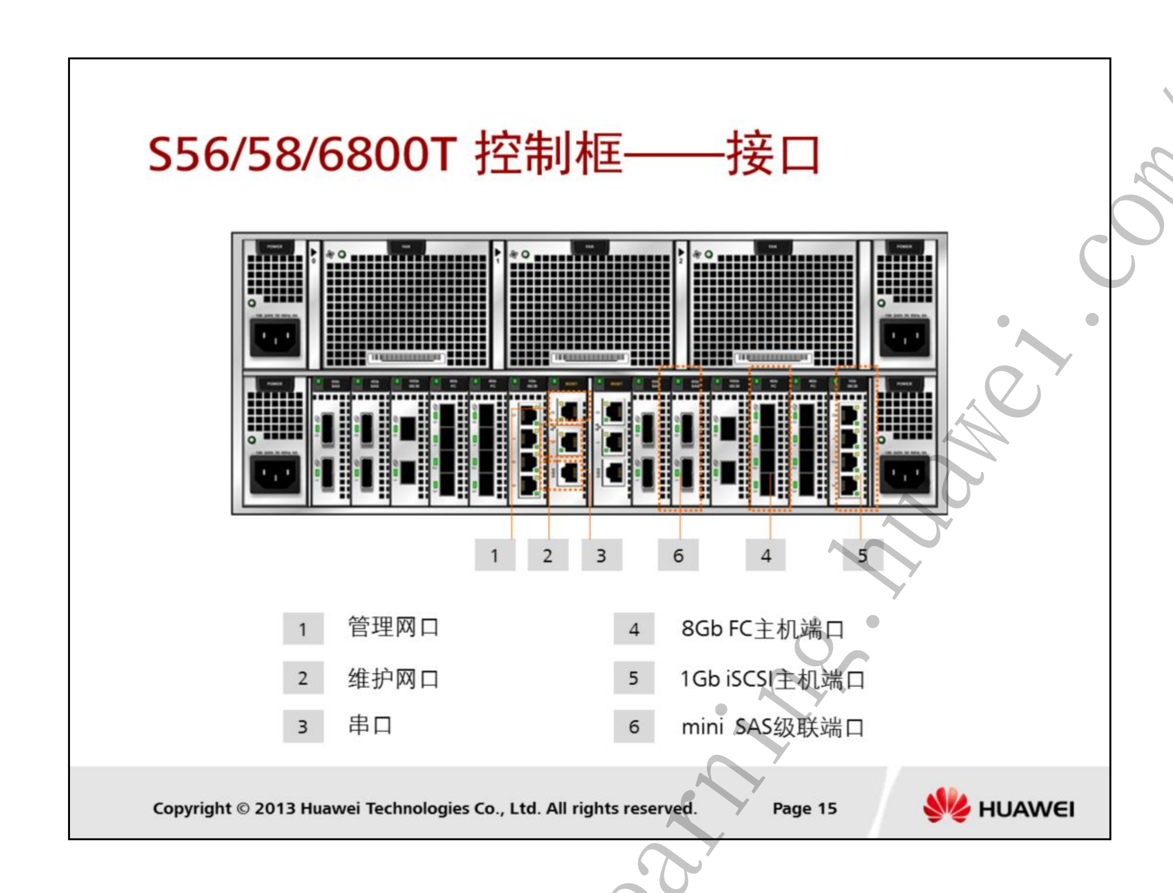


S5600T/5800T/6800T为4U控制框，主要由控制器模块、BBU模块、电源模块、风扇模块以及多个接口模块组成，是存储系统的核心部件。

控制框可使用交流电源模块或直流电源模块进行供电。



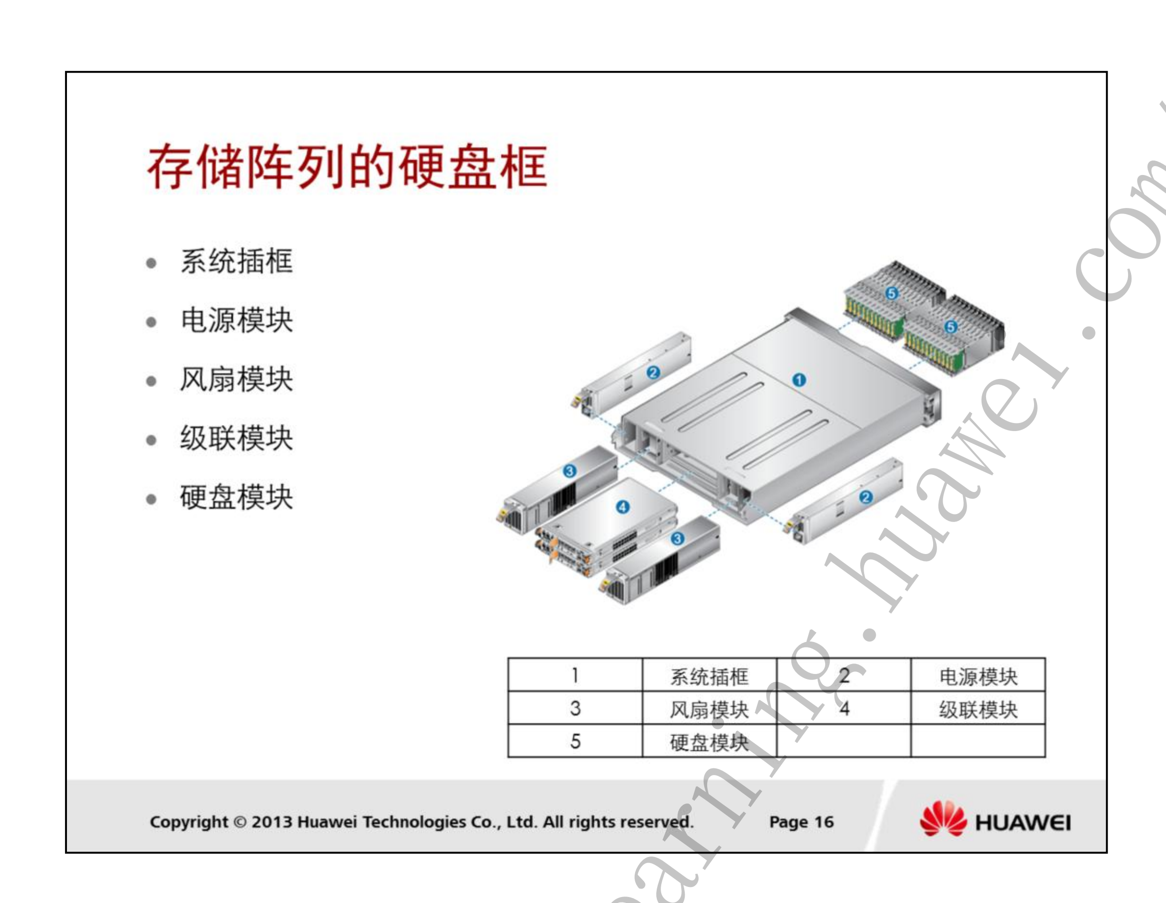
由于S5600T/S5800T/S6800T是盘控分离架构，其控制框是不带硬盘的。若掉电后仍将Cache脏数据刷入硬盘中，那就必须要同时对控制框以及硬盘框中的硬盘进行供电，这就意味着硬盘框也需要一个专门的UPS。因此，S5600T/S6800T在控制器中内置一块SSD，掉电后Cache脏数据会写入这块SSD盘中，控制框的内置BBU可以对内置SSD供电。而系统配置数据、系统日志等信息仍然保存在硬盘框的前4块硬盘中，但这部分数据是不需要进行掉电保护的。



管理模块为存储系统提供管理接口，主要包括管理网口、维护网口和串口。管理模块将系统配置数据、告警信息以及日志信息保存到保险箱硬盘上。

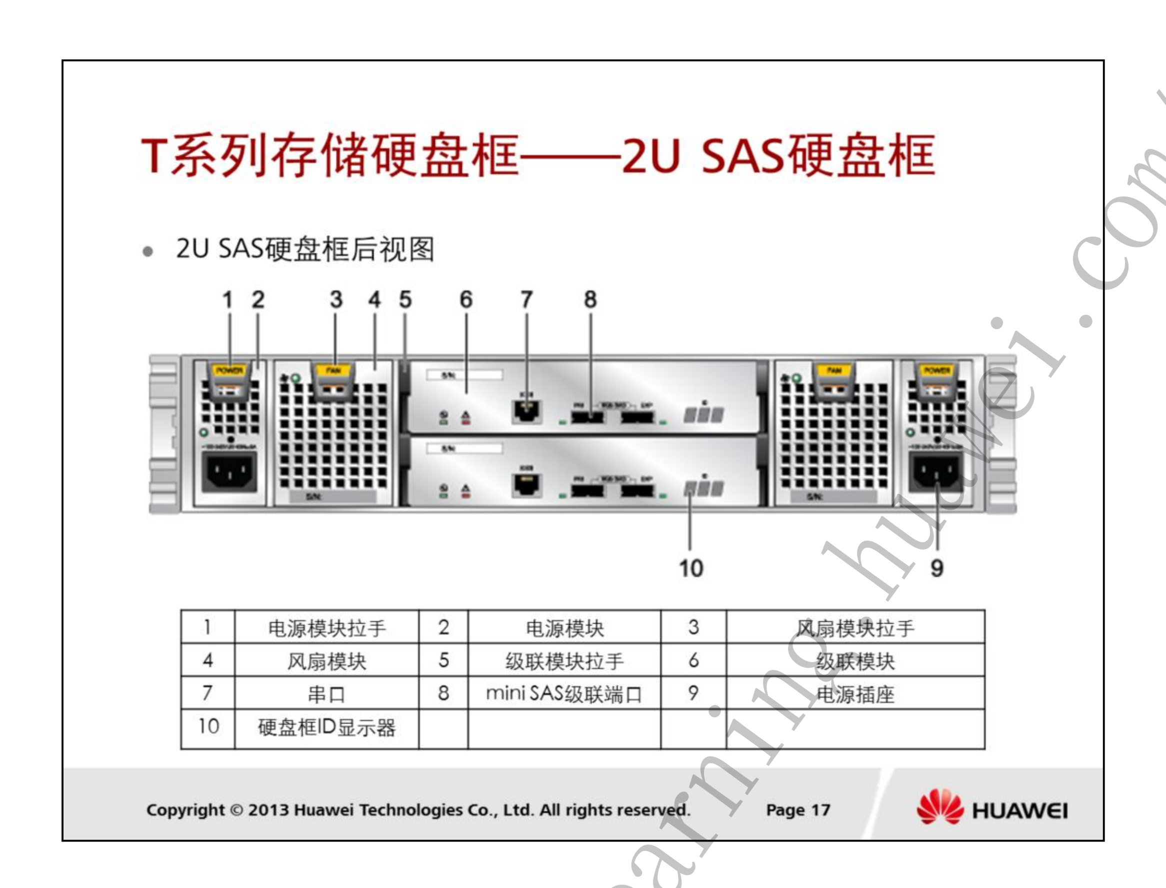
I/O接口模块主要用于存储设备与应用服务器之间的连接，存储设备支持的I/O接口模块包括8Gb FC接口模块、10Gb TOE接口模块、1Gb iSCSI接口模块以及10Gb FCoE接口模块。

级联模块通过级联端口连接控制框和硬盘框，是控制框和硬盘框之间进行数据传输的连接点。 SAS级联模块提供2个传输速率为6Gbit/s的mini SAS级联端口，用于级联硬盘框。控制框的SAS级联模块通过mini SAS电缆与存储系统的后端硬盘阵列连接。当连接的设备传输速率低于级联端口速率时，级联端口将自动适应传输速率，以保证数据传输通道的连通性和数据传输速率的一致性。



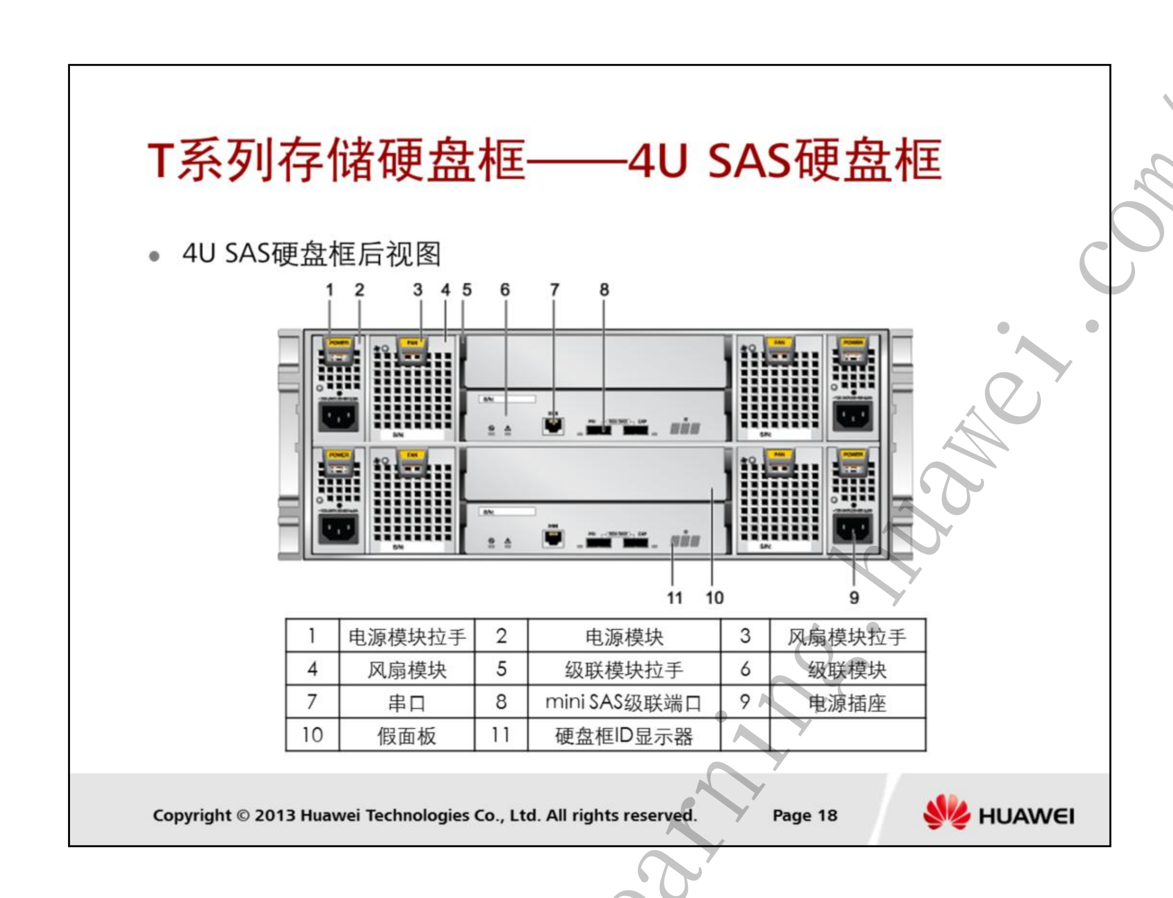
1	系统插框	2	电源模块
3	风扇模块	4	级联模块
5	硬盘模块		

硬盘框由系统插框、级联模块、硬盘模块、电源模块以及风扇模块等组成，是存储空间的载体。



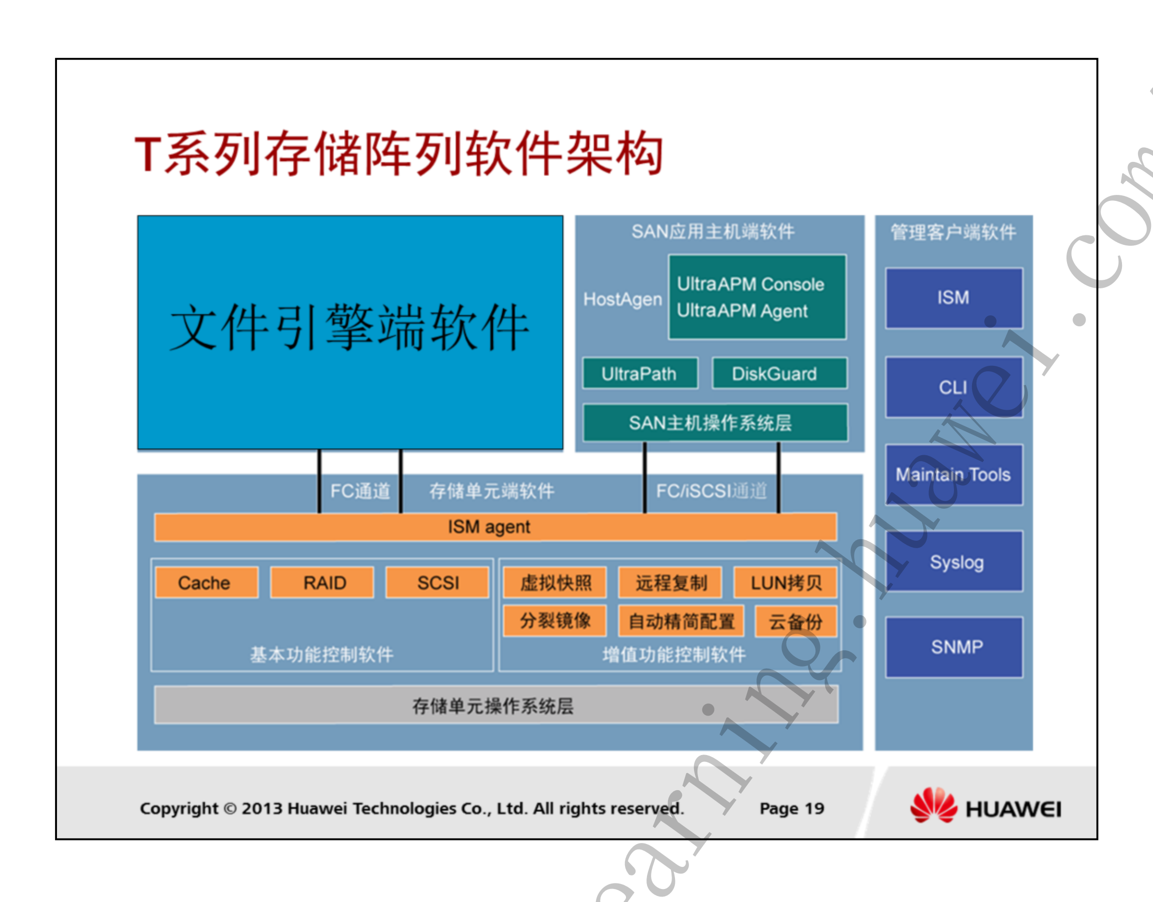
硬盘框由系统插框、级联模块、硬盘模块、电源模块以及风扇模块组成，是存储空间的载体。

2U SAS硬盘框可满配24块2.5英寸硬盘，同时可使用交流电源模块或直流电源模块进行供电，用户可以根据应用环境选择相应的电源模块。



1	电源模块拉手	2	电源模块	3	风扇模块拉手
4	风扇模块	5	级联模块拉手	6	级联模块
7	串口	8	mini SAS级联端口	9	电源插座
10	假面板	11	硬盘框ID显示器		

4U SAS硬盘框可满配24块3.5英寸硬盘，可使用交流电源模块或直流电源模块进行供电，用户可以根据应用环境选择相应的电源模块。



存储系统提供丰富的全套存储管理软件，方便您轻松快捷地管理和维护存储系统。存储系统软件由文件引擎端软件、存储单元端软件、SAN应用主机端软件和管理客户端软件4部分组成。

- 文件引擎端软件

通过文件系统卷管理、文件系统快照等模块以卷的方式将硬盘或虚拟的LUN进行虚拟化管理以及基于文件系统的快照等功能。这部分内容将在其他高级课程中介绍，本课程不做深入介绍。

- 存储控制软件

实现主机与存储系统的通信和数据交互，实现存储系统的硬件监控、存储系统基本功能、增值功能的管理和配置。每一项功能的具体内容将在后面章节中介绍。

- SAN应用主机端软件

通过UltraPath实现应用服务器到存储设备之间的路径选择。UltraAPM为应用系统提供数据一致性保障，并提供针对应用服务器、应用系统、存储系统进行数据保护与容灾任务的配置、管理、监控等工作。

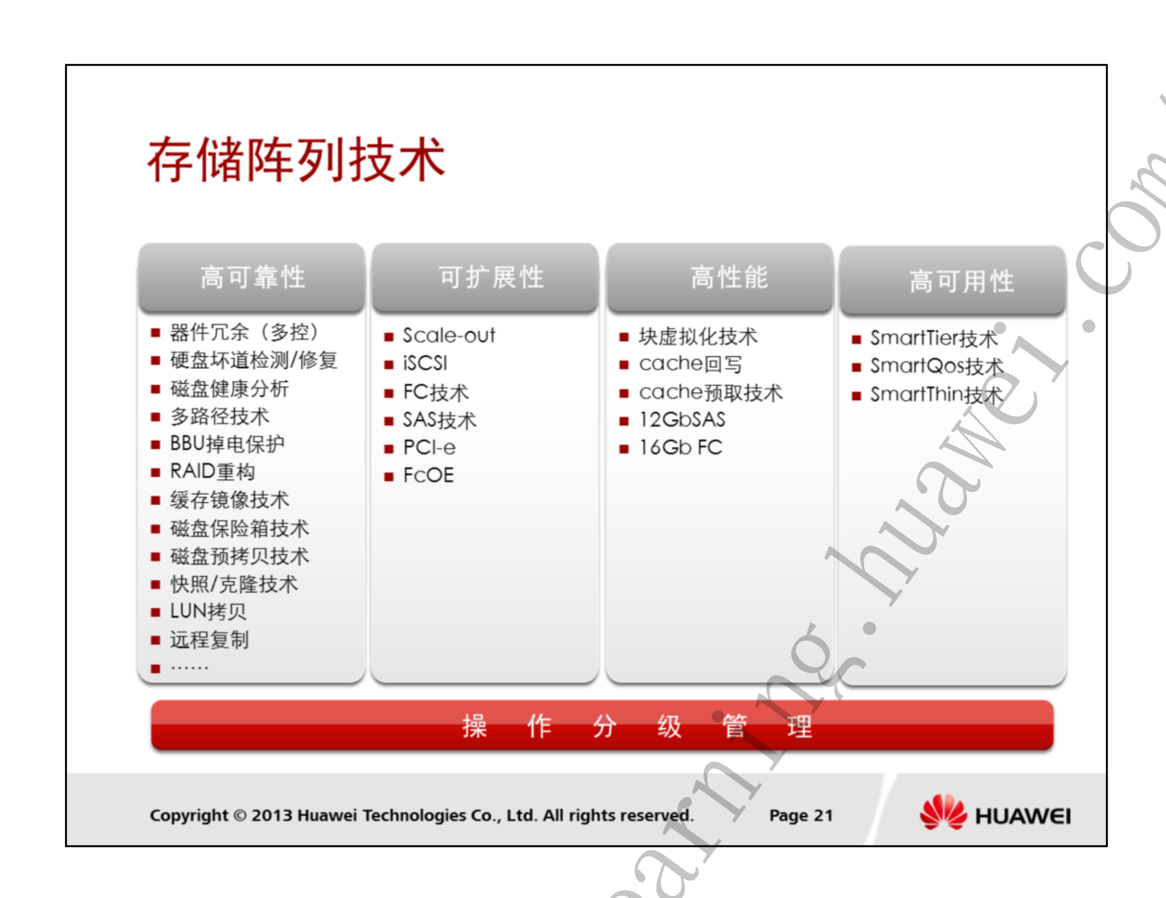
- ISM管理客户端软件：
  - ISM agent集成在存储控制软件中，是部署在存储设备上的ISM通信接口，实现数据收集
  - ISM client部署在管理客户端，用于对存储系统进行配置、管理和监控。
  - ISM server是实现ISM内部逻辑处理、数据组织和协议组装解析等。



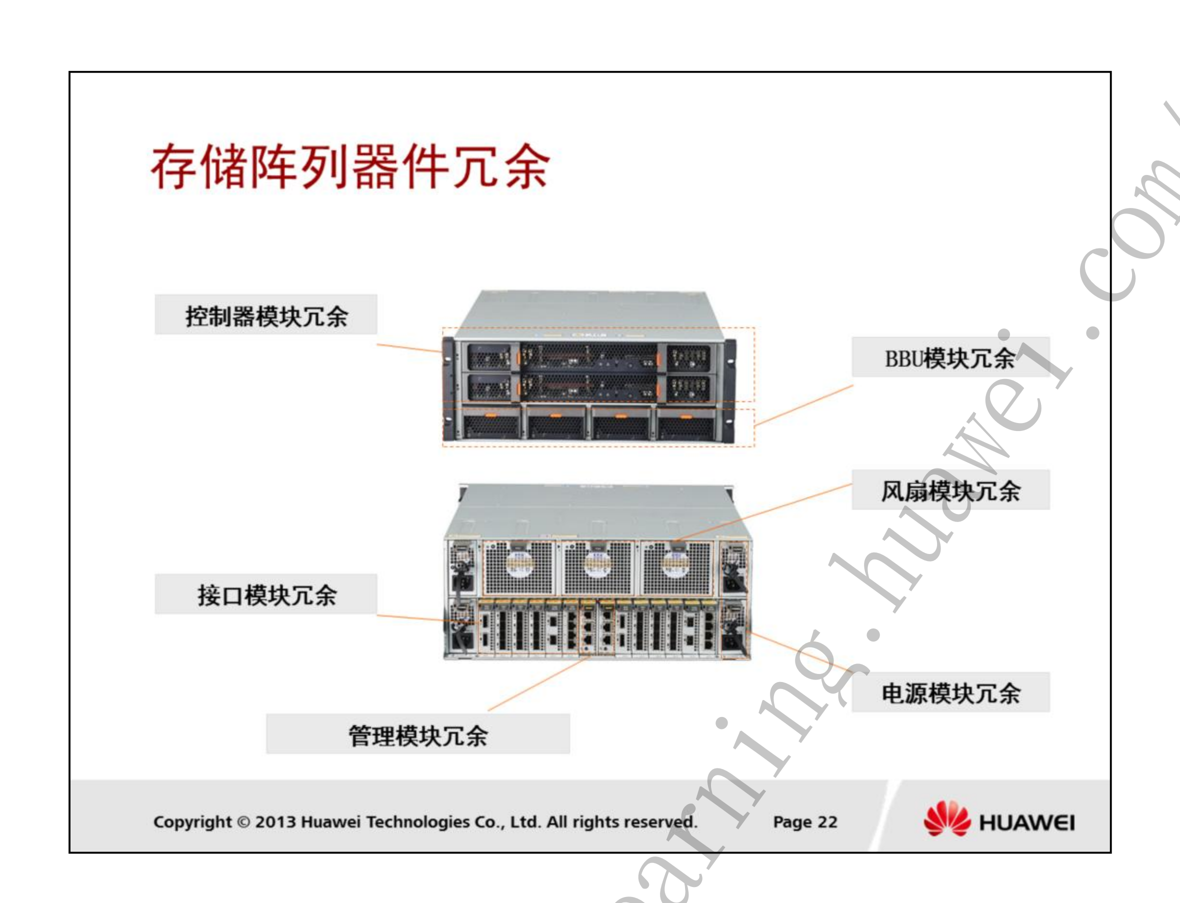
目录

 Page 20





存储阵列为保证数据存取业务的高可靠性、可扩展性、高性能、高可用性，开发和使用了相应的技术，下面我们对相关技术做简要介绍，在后续的存储课程中，还将进行详细介绍。



存储阵列系统实现了控制器、管理模块、BBU模块、接口模块、电源模块、风扇模块等部件的冗余，存储系统的可靠性得到保障。同时，因为采用双控双活技术，大大提升了存储系统的数据存取效率。

存储阵列的多控技术

- 双控制器系统的工作模式:
  - 主备模式
  - 双活模式

- 在双控制器系统中，有以下两类典型的工作模式；
  - Active-Passive工作模式：又被称为主备模式，即任何时刻两个控制器中只有一个控制器处于激活状态，作为主控制器，用于处理应用服务器的I/O请求，而另外一个控制器处于空闲状态，作为备用控制器，以便在主控制器出现故障或者处于离线状态时及时接管其工作。
  - Active-Active工作模式：常被称为双活模式，顾名思义即两个控制器都处于激活状态，可并行处理来自应用服务器的I/O请求，一旦某个控制器出现故障或离线，另一个控制器将及时接管其工作，且不影响自己现有的任务。可见，这种工作模式通过互为冗余备份来确保系统高可靠性的同时，它还具有均衡业务量、充分利用资源，提升系统性能等诸多优点。

硬盘坏道检测技术



面对磁盘坏道，被动应对还是主动出击?

读写失败自动分析

- 读写失败有多重原因
- 根据系统当前状态、硬盘当前状态、IO失败信息等进行综合分析



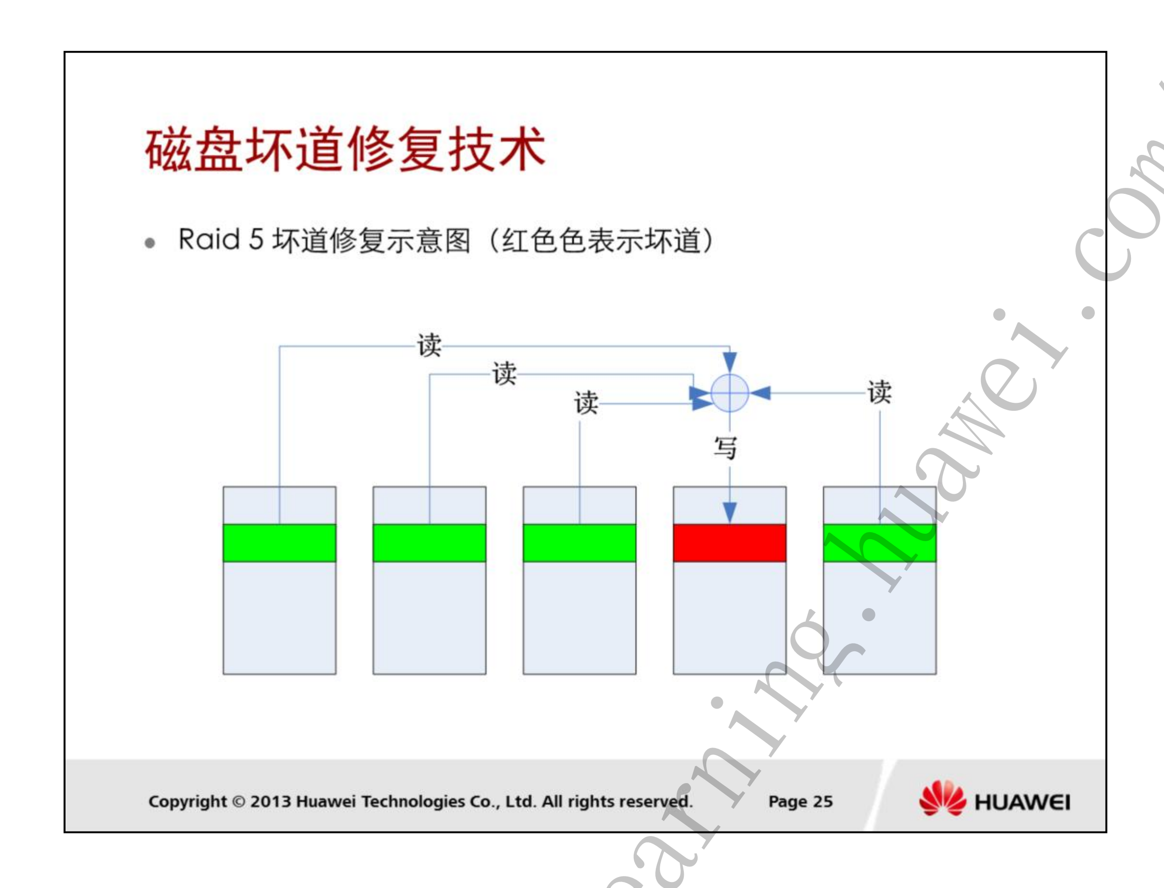
- 直接使用硬盘的内建介质扫描功能
- 避免了硬盘扫描对后端带宽的占用
- 将对系统性能的影响降到最低

 Page 24 HUAWEI

据统计，SATA硬盘年故障率约为2.5%，其中70%为可修复坏道。以30天为周期对硬盘进行周期性扫描，就意味着硬盘坏道每30天被全部发现并修复一次。由此，可以计算出实现硬盘介质扫描技术之后，硬盘年故障率可降低到$\lambda=2.5\%\times0.3+2.5\%\times0.7\times30/365=0.89\%$，Raid组失效率可以降低约1个数量级。

可以通过两种途径来检测硬盘坏道。

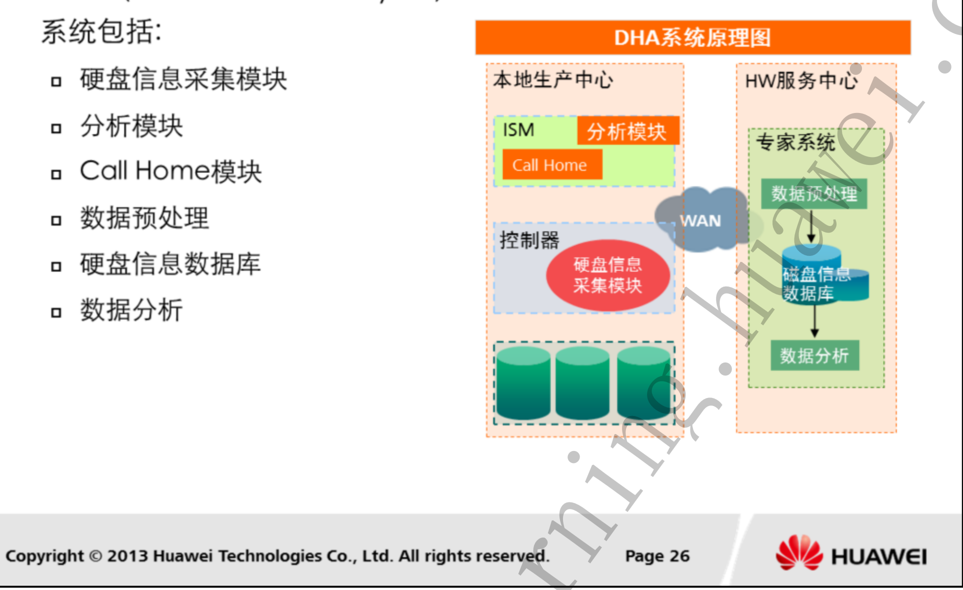
- 读写失败自动分析。硬盘读写失败可能有多种原因，如硬盘磁头损坏、硬盘接口损坏、连接线缆损坏、硬盘被拔出等。当硬盘读写失败发生时，存储系统会根据系统当前状态、硬盘当前状态、IO失败信息等进行综合分析，判断该次读写失败原因是否为硬盘坏道。
- 硬盘介质主动扫描。存储系统还支持硬盘介质后台扫描技术，利用硬盘空闲周期检查硬盘介质，及时发现硬盘坏道，避免累积错误。存储系统的硬盘介质扫描，摈弃了顺序读取硬盘所有扇区的传统方式，直接使用硬盘的内建介质扫描功能，避免了硬盘扫描对后端带宽的占用，将对系统性能的影响降到最低。当对正在进行介质扫描的硬盘进行读写时，扫描动作自动停止，转而处理读写操作，在读写停止之后，硬盘可以从之前的断点继续扫描。



一直以来，硬盘都是计算机系统中最主要的存储设备，同时也是计算机系统中最容易出故障的部件。众所周知，常见的硬盘都是采用磁介质来进行数据的存储，满足计算机系统IO处理的需要。硬盘的出货量大大增加，同时也意味着硬盘报废的数量大大增加。存储系统支持硬盘坏道智能修复技术，通过该技术能够对硬盘坏道进行智能修复，使硬盘恢复健康，有效延长硬盘寿命，从而减少硬盘报废数量，降低客户TCO。

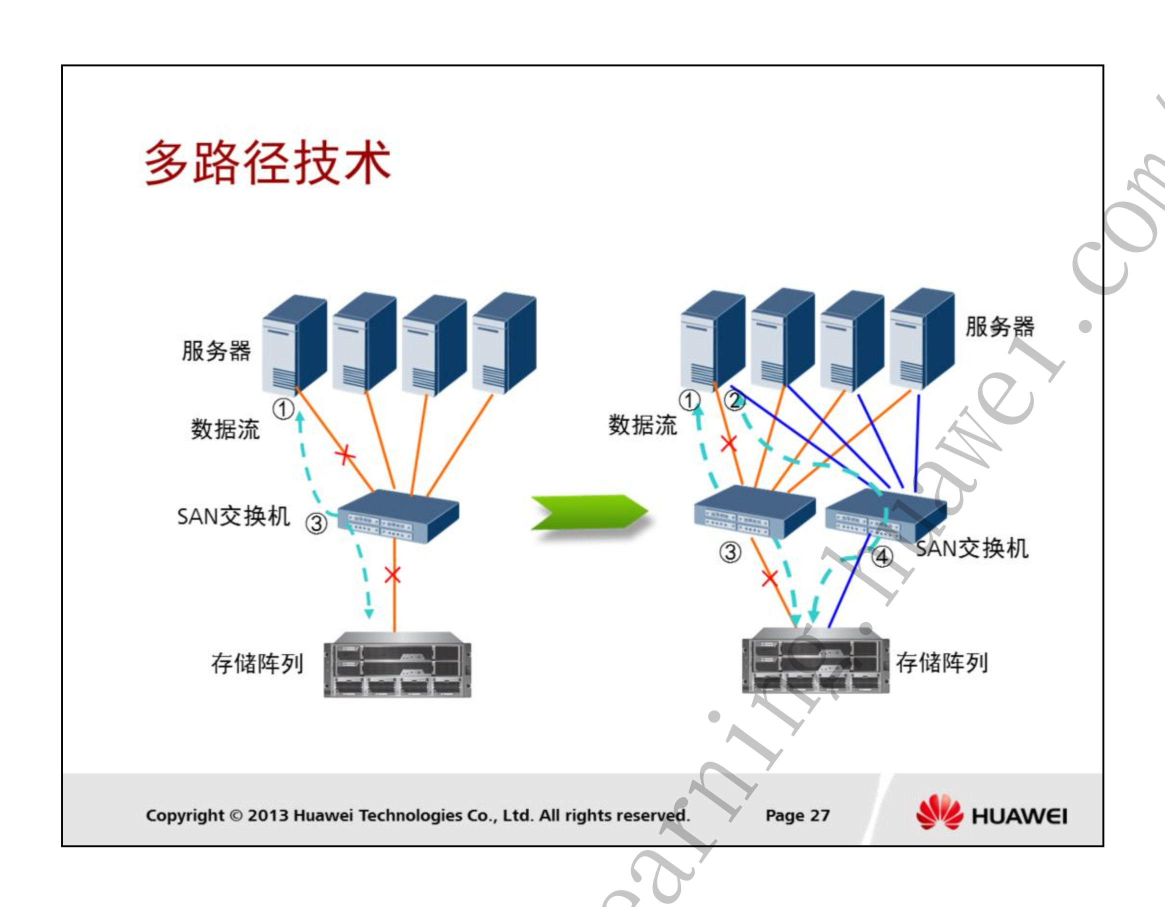
存储系统的修复策略如下：

- 如果发现坏道的硬盘为空闲盘或空闲热备盘，则直接使用全零数据对坏道地址进行覆盖写操作，以触发硬盘的坏道地址重映射操作，完成硬盘坏道修复。
- 如果发现坏道的硬盘为RAID组成员盘或已用热备盘，则不能使用全零数据对坏道进行覆盖写，因为该地址可能已包含有效的用户数据，此时需要先恢复出该地址的正确数据，然后使用恢复出的正确数据重写该坏道地址，完成坏道修复。
- 在使用上述策略2的情况下，需要通过RAID组的冗余来恢复重写坏道地址所需正确数据（RAID0无冗余，无法进行恢复）。只要RAID5中没有两个盘同一地址坏道，RAID10镜像组中的互为镜像的两个盘没有同一地址坏道，RAID6没有三个盘同一地址坏道，存储系统都可以对坏道进行成功修复。



DHA（Disk Health Analyzer）系统的主要目的是采集网上硬盘的运行状态信息，并将信息回传至服务中心，以建立硬盘信息数据库，及硬盘健康度分析模型，通过该模型对现网硬盘的健康状况进行诊断。

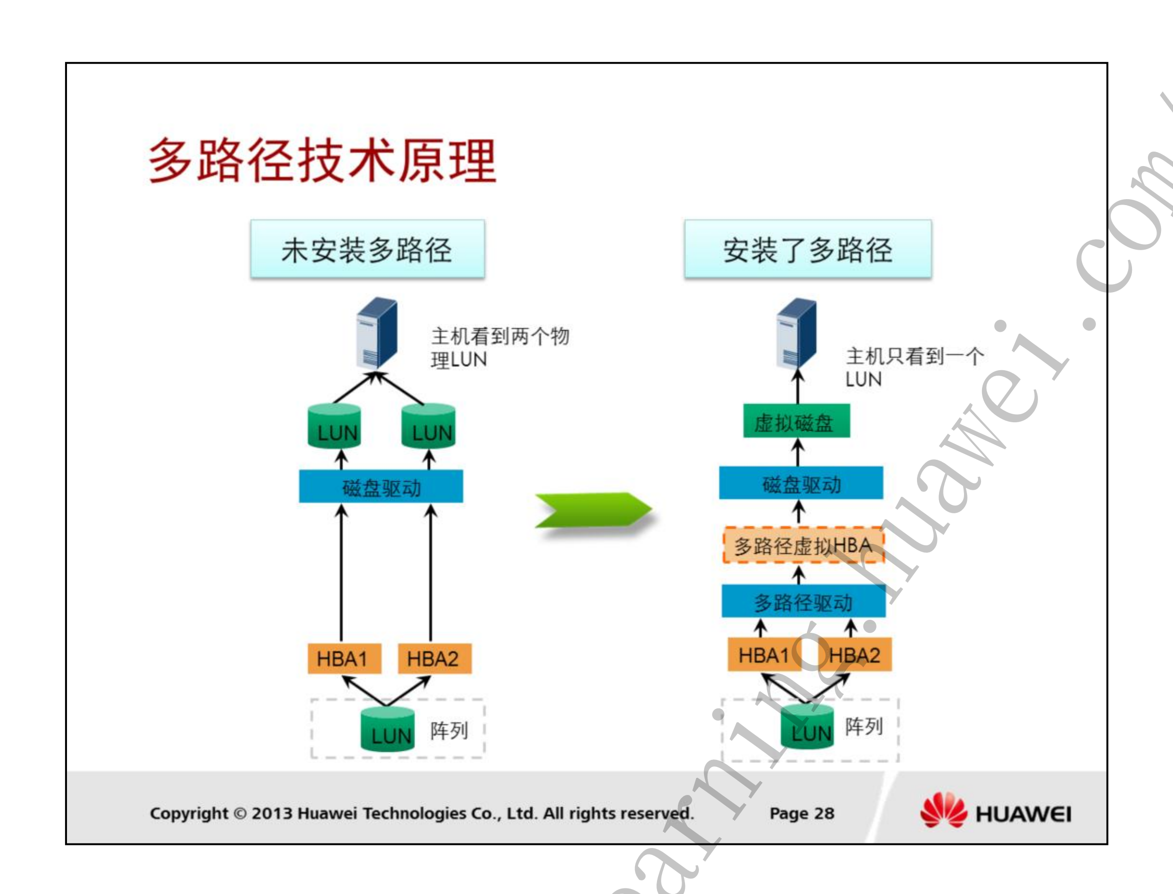
- 实现过程及原理
  - 阵列控制器中集成了DHA系统的信息采集模块，硬盘信息由该模块收集后上传给ISM
  - ISM中集成了DHA系统的分析模块，用于分析收集到的硬盘信息并作出诊断，并通过Call Home功能，将这些信息通过网络上传给HW服务中心。
  - DHA系统的功能可通过ISM中的开关手动打开或关闭。



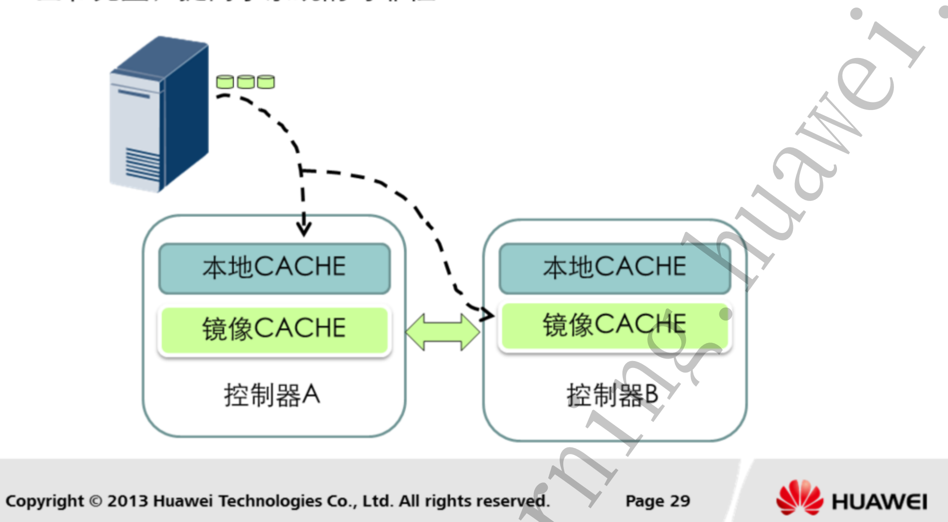
用户数据从主机侧到磁盘阵列，经历的典型路径为：主机 -> SAN网络 -> 存储系统机头 -> 存储系统磁盘。

所谓多路径技术，即在一台主机和存储阵列端使用多条路径连接，使主机到阵列的可见路径大于一条，其间可以跨过多个交换机，避免在交换机处形成单点故障。如上右图中，主机到存储阵列的可见路径有2条，即（1，3）（2，4），该路径上有两台独立的交换机。在这种模式下，当路径1断开时，数据流会在主机多路径软件的导引下选择路径（2，4）到达存储阵列侧，同样在左侧交换机失效时，也会自动导引到右侧交换机到达存储阵列。在路径1恢复的情况下，IO流会自动切回原有路径下发。整个切换和恢复过程对主机应用透明，完全避免了由于主机和阵列间的路径故障导致IO中断。

存储系统冗余保护方案涉及了这个路径上的所有领域，在主机侧和SAN网络领域，通过结合UltraPath多路径软件及其它多路径软件，保证了前端路径没有单点故障；在存储机头侧，使用了全冗余硬件及热插拔技术实现了双控双活的冗余保护；在磁盘侧，利用磁盘双端口技术及磁盘多路径技术，实现了磁盘侧冗余保护。



- 多路径软件的作用：
  - 避免了同一LUN有多条路径可达导致的操作系统逻辑错误
  - 增强了链路的可靠性，避免了因为单个链路故障而导致的系统故障



Cache即高速缓存，是存储系统中用于缓存主机数据，提高阵列性能及可靠性的部件。Cache主要用来缓存主机数据，使主机在进行数据读写时不用每次都对硬盘进行操作，从而提高系统的读写性能。同时存储阵列通常使用双主控，且采用双控双活技术，双控制器都存在相同大小相同规格的Cache，使得数据在本端控制器Cache和对端控制器Cache都有缓存，即两个控制器的写Cache数据通过相互镜像实现备份，确保数据的安全和完整，提高了系统的可靠性。

Cache镜像技术实现

- 镜像通道
  - SAS
  - PCI-E
  - FC
- 实现机制
  - 读Cache
  - 写Cache
  - 镜像Cache

镜像通道
控制器 A
控制器 B

互为镜像
镜像通道
读 Cache
写 Cache
镜像 Cache
Date1
Date2

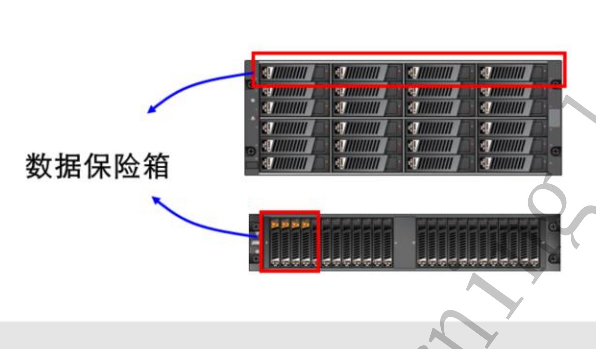
 Page 30 HUAWEI

Cache镜像技术是通过镜像通道实现的。可采用SAS、PCI-E或FC作为镜像通道。由于在物理上采用了两条镜像通道，使系统在可靠性上得到了进一步的提升，即使其中一条通道出现故障也不会影响业务，保证业务的连续性。

- 实现机制：
  - 读Cache：作用为缓存读数据；在有主机读请求下发时，Cache会先将磁盘中数据读到内存中，并启动预取，根据数据的连续性将磁盘中后续可能会被主机读到的数据预取到Cache中，增加后续主机请求的读命中概率。
  - 写Cache：作用为缓存写数据，在有主机写请求下发时，Cache会先将数据写到内存的固定区域中，并返回给主机写完成，后续Cache会将这部分数据整合后统一写到硬盘中，借此提高主机写的效率。
  - 镜像Cache：作用为缓存写数据并在对端控制器Cache保存这部分数据的镜像，在有主机镜像写请求下发时，Cache会将这部分数据分别写到本端控制器的内存及对端控制器的内存，保证两端数据的一致性。实际上，本端的镜像Cache就是对端写Cache的镜像，是对端写Cache的备份。这样的结构充分保证了系统的高可用性。

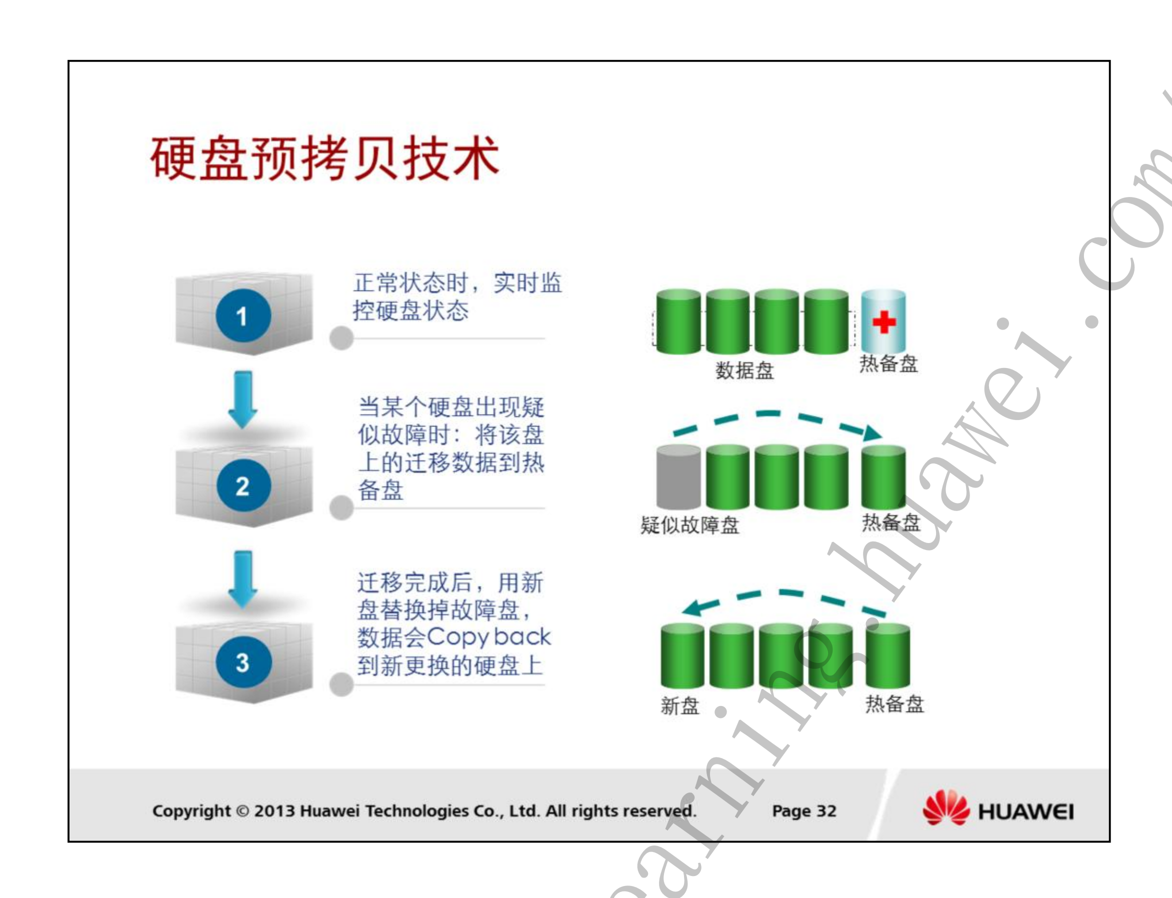
数据保险箱技术

- 数据保险箱技术用于保存Cache数据，避免因系统意外断电时数据丢失。
- 内置BBU电池可保证在系统意外断电时，对Cache和系统保险箱硬盘同时供电，让Cache中的数据写到数据保险箱中，实现Cache数据永久保存。

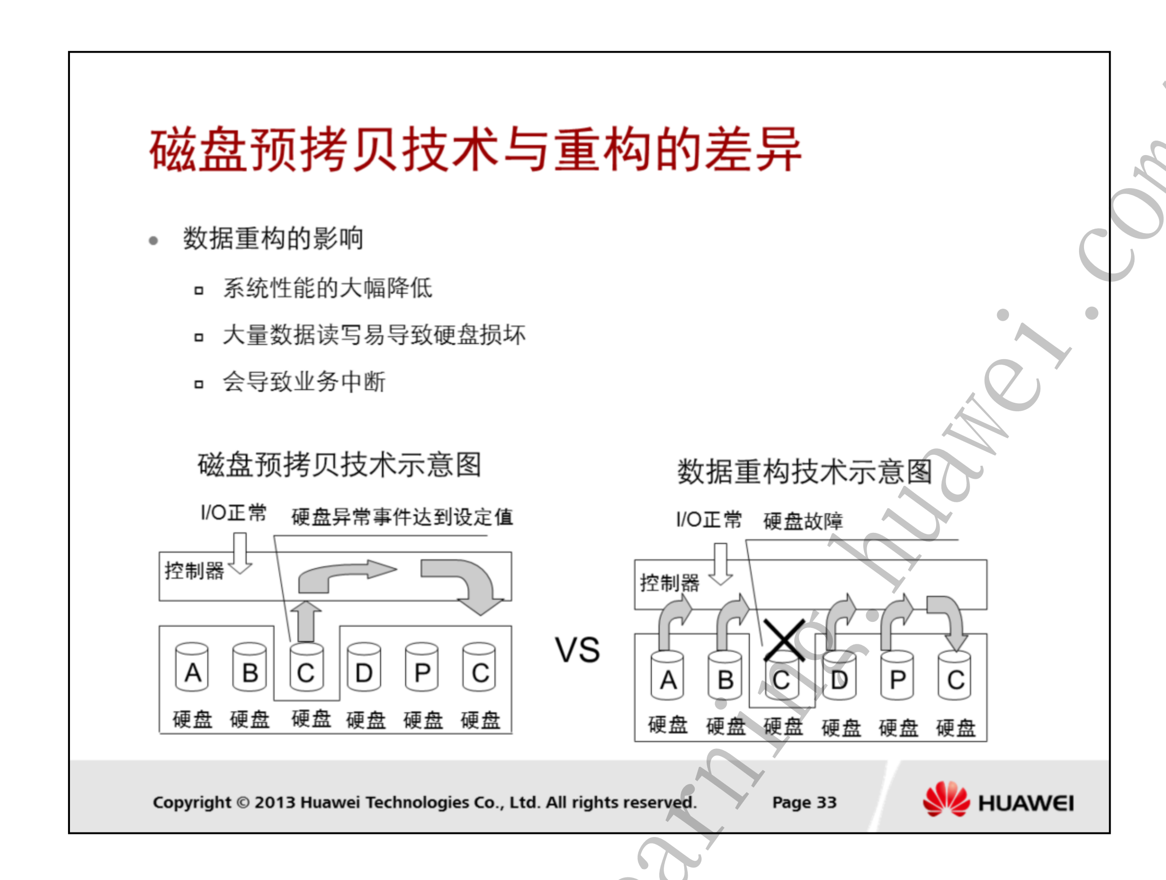




数据保险箱由控制框前4块盘的最后部分空间组成。保险箱硬盘空间之间使用RAID10保护，每块盘除系统占用的最后部分空间外，剩余的绝大部分空间均可分配给用户做正常的数据存储使用。

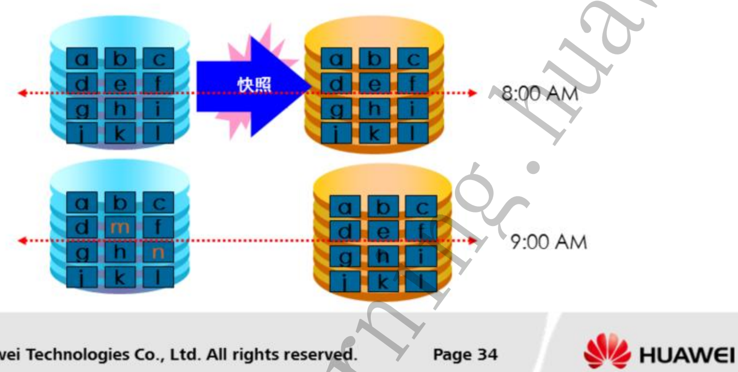


- 技术原理
  - 系统实时从硬盘的SMART信息中读取硬盘的状态信息，当发现硬盘错误统计超过设定阀值后，立即启动将数据从疑似故障盘中迁移到热备盘，同时向管理人员告警，提醒更换疑似故障盘。
- 技术优势
  - 大大降低重构事件发生的概率
  - 提高系统的可靠性



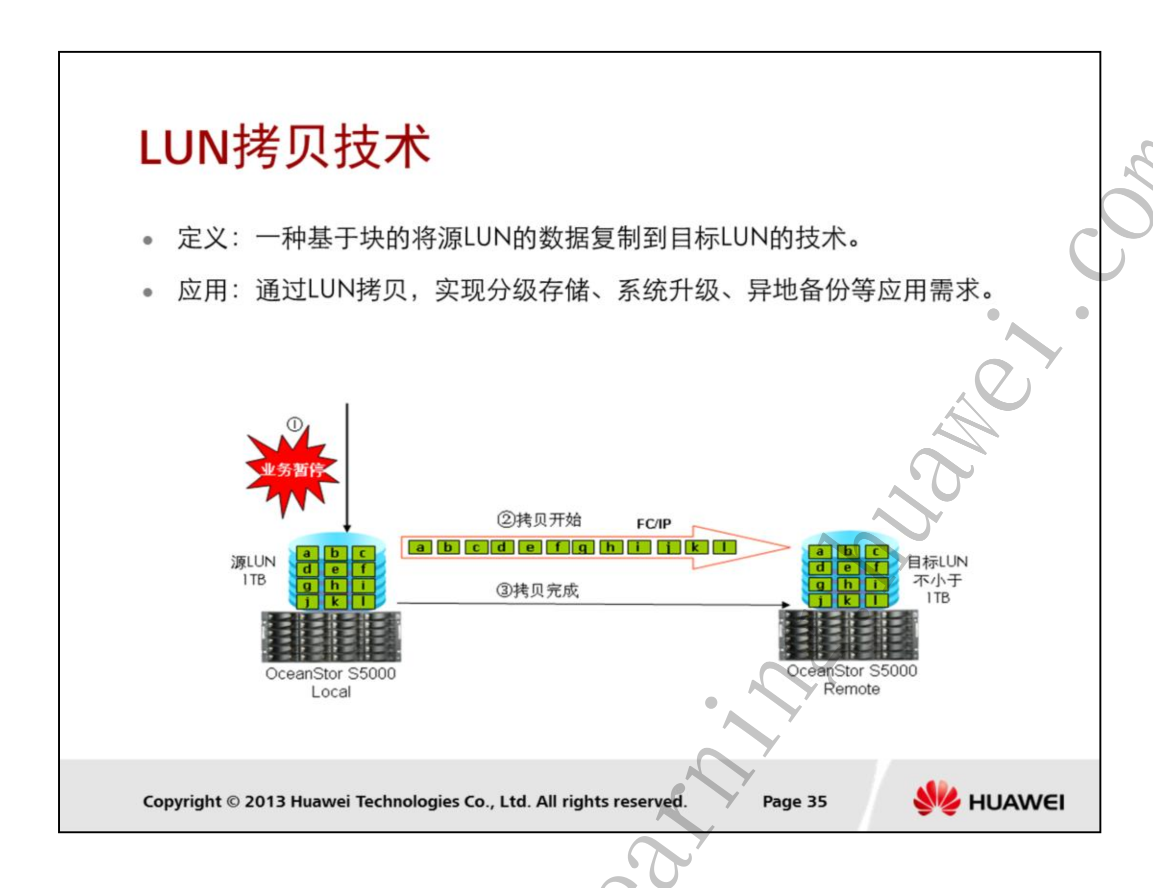
磁盘预拷贝技术可以充分利用从检测到即将失效到磁盘真正失效这段时间来降低风险，与数据重构技术相比，磁盘预拷贝技术具备以下优势：

- 低风险：如图所示，整个预拷贝过程中，RAID组处于正常状态，所有成员盘均处于可用状态，RAID组的数据冗余信息是完整的，客户数据无丢失风险。而在重构过程中，RAID组处于降级状态，RAID组的数据冗余信息不完整（或已丢失），客户数据处于高风险状态。
- 高效率：如图所示，重构过程中要涉及RAID中的多个盘，速度较低，而且占用后端带宽也较大。而磁盘预拷贝技术只是两个硬盘之间的数据拷贝，速度快，占用资源少（和重构相比）。



快照不做完整的物理上的数据拷贝，而是通过写前拷贝技术和映射表得到源数据在某一时间点的数据副本。

- 快照的目的
  - 快照可以作为备份和归档的数据源。
  - 快照可以对存储设备上的数据灵活且频繁地生成多个时间点的数据副本，在需要时可以快速地恢复数据。
- 快照可以给用户的数据备份带来如下受益：
  - 可以瞬间生成，不影响主机业务，得到源LUN在某个时间点的数据副本。
  - 可以根据用户自定义的备份策略，对存储设备上的数据灵活且频繁地生成多个时间点的数据副本，从而方便快速恢复数据。



- LUN拷贝特点：
  - 将数据从本存储系统复制到一个或多个其他存储系统。
  - 将数据从一个或多个其他存储系统复制到本存储系统。
  - 将数据从本存储系统中的一个LUN复制到另一个LUN。
- LUN拷贝的优势：
  - 高性能，LUN拷贝的实现过程比基于主机拷贝的实现过程更加简洁，因此LUN拷贝的性能将大大高于基于主机的拷贝。
  - 业务无关性，LUN拷贝的执行过程中，不需要主机参与，不会占用主机的资源。主机可以将更多的资源用于处理生产业务。
  - 高兼容性，LUN拷贝能够在异构环境下部署，支持不同品牌存储系统间的LUN拷贝。
- LUN拷贝支持全量和增量两种拷贝模式。

远程复制（Hypermirror）技术

- HyperMirror(远程复制)提供不同区域间数据的同步/异步镜像
- 保护用户数据，避免灾难性事件带来的损失。

主站点 主LUN 从LUN 从站点

 Page 36 HUAWEI

远程复制，是数据镜像技术的一种，它能够在两个或多个站点维护若干个数据副本，利用长距离来避免灾难发生时的数据丢失。

实现远程镜像功能的技术较多，业界使用最为广泛的主要有同步远程镜像和异步远程镜像两种。同步远程复制常用于距离较近网络延迟比较小的场景中。异步远程复制常用于距离较远网络延迟稍微大的场景中。

- 远程复制技术的优势
  - 快速业务切换
  - 全冗余的IO传输路径
  - 自动恢复策略
  - FC/IP网络支持
  - 一致性组
  - 增量同步技术
  - 复制速率动态调整
  - 全面的管理手段

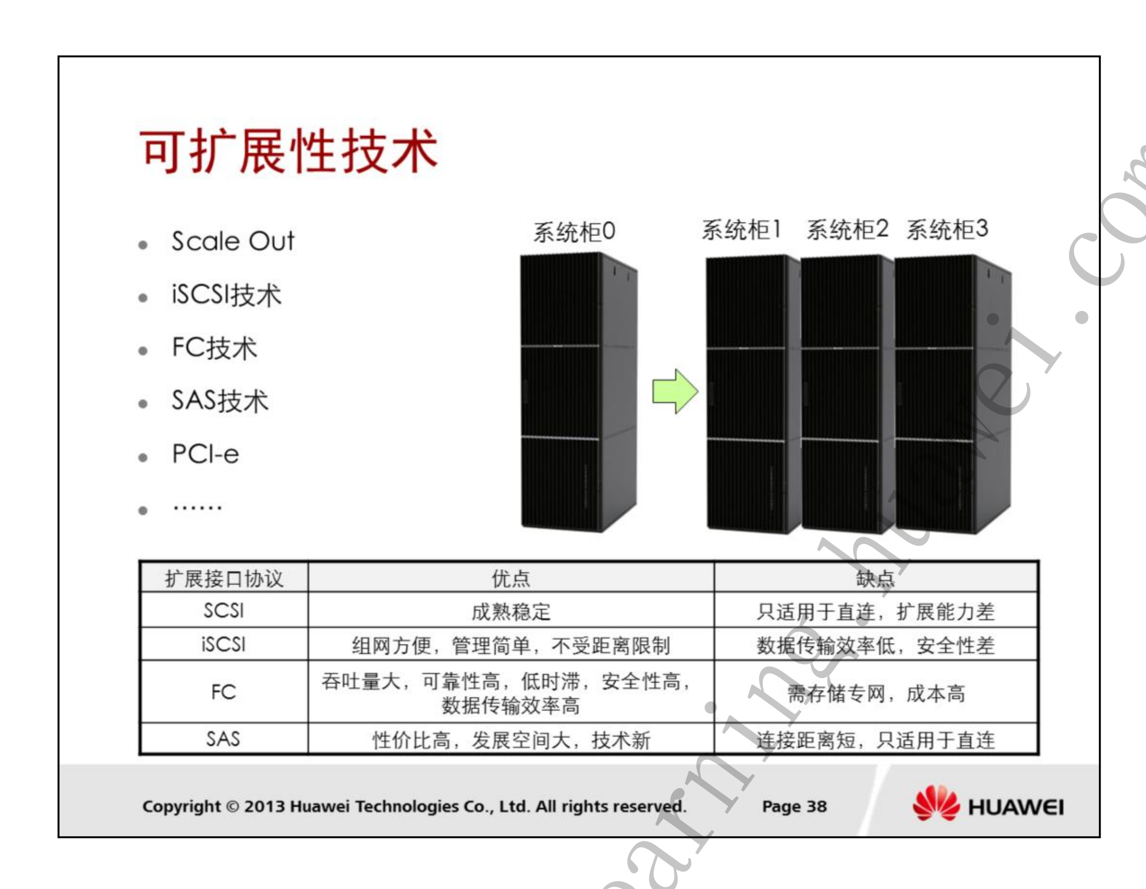
LUN拷贝与远程复制的区别

对比项目	远程复制	LUN拷贝
兼容性	只能在同类型存储系统之间运行。	不但支持同类型存储系统，而且支持经过认证的第三方的存储系统。
数据下发	每个主LUN只能向1个（异步模式）或2个（同步模式，分别位于不同存储系统上）从LUN复制数据。	每个源LUN可以向数十个或者更多目标LUN复制数据。
数据备份	用于持续的数据保护。从LUN可读，但始终不可写。	用于数据备份。数据拷贝完成后，主机即可访问目标LUN。

 Page 37



虽然远程复制也能够将数据从本存储系统复制到另一个存储系统，但是在具体实现方面与LUN拷贝相比还是有很大的差异。



扩展接口协议	优点	缺点
SCSI	成熟稳定	只适用于直连，扩展能力差
iSCSI	组网方便，管理简单，不受距离限制	数据传输效率低，安全性差
FC	吞吐量大，可靠性高，低时滞，安全性高，数据传输效率高	需存储专网，成本高
SAS	性价比高，发展空间大，技术新	连接距离短，只适用于直连

- Scale out：通过一定的分布式算法将一个个独立的低成本存储节点组成一个大而强的系统，即向外扩展。
- SCSI（Small Computer System Interface，小型计算机系统接口）：最初是一种为小型机研制的接口技术。 SCSI成为一种连结主机和外围设备的接口，支持包括磁盘驱动器、磁带机、光驱、扫描仪在内的多种设备，主要功能是在主机和存储设备之间传送命令、状态和块数据。
- iSCSI（Internet SCSI）：把SCSI命令和块状数据封装在TCP中在IP网络中传输，基本出发点是利用成熟的IP网络技术来实现和延伸SAN。
- FC（Fibre Channel）即光纤通道，是一种用于构造高性能信息传输的、双向的、点对点的串行数据通道，主要应用于FC SAN存储网络中。
- SAS(串行连接SCSI):是一种访问计算机外围设备的方法，它在电缆中使用串行方式传输数字数据。 串行连接SCSI优于早期的并行技术，串行数据传输允许使用比并行数据传输更长的电缆。
- PCI-e：即PCI Express，是新一代的总线接口。采用了目前业内流行的点对点串行连接，比起PCI以及更早期的计算机总线的共享并行架构，每个设备都有自己的专用连接，不需要向整个总线请求带宽，而且可以把数据传输率提高到一个很高的频率，达到PCI所不能提供的高带宽。

以上相关技术在后面的章节中将有更详细的描述。

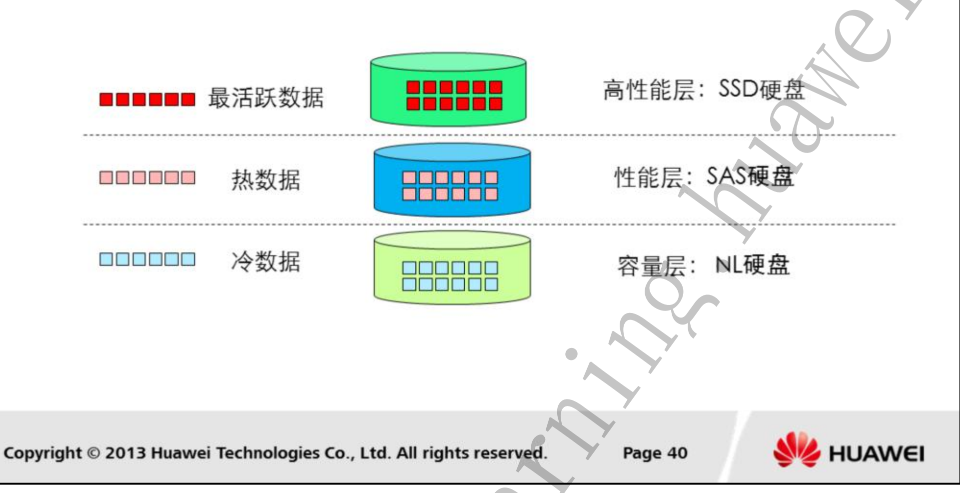
块级虚拟化技术

- 一种新型RAID技术。该技术将硬盘划分成若干固定大小的块（chunk），然后将其组合成若干个小RAID组。

Page 39
HUAWEI

块虚拟化是一种新型RAID技术。将硬盘划分成若干固定大小的块（chunk），然后将其组合成小RAID组（CKG）。RAID的组成不再以硬盘为单位，而以chunk为单位。

- 块虚拟化技术的特点：
  - 将数据分布到系统中所有硬盘，充分发挥系统的读写处理能力。
  - 某一硬盘失效时，存储池内的其它硬盘都会参与重构，消除传统RAID下的重构性能瓶颈，提高了重构数据的速度。
  - 存储系统按照用户设置的“数据迁移粒度”将CKG划分为更小的extent，若干extent组成了用户需要使用的LUN。在存储系统中申请空间、释放空间、迁移数据都是以extent为单位进行的。



- 存储系统支持的存储介质包括：
  - SSD（Solid State Drive）硬盘
  - SAS（Serial Attached SCSI）硬盘
  - NL（Near Line） SAS硬盘

不同存储介质在存储成本和存储性能方面的差异很大，这导致用户难以在存储成本和存储性能之间权衡。SSD硬盘的响应时间很短，每单位存储请求处理成本很低，但每单位存储容量成本很高。NL SAS硬盘每单位存储容量成本较低，但响应时间很长且每单位存储请求处理成本很高。SAS硬盘介于以上两者之间。

SmartTier能提高存储系统整体性能并降低用户成本。成本包括硬件和软件的采购、空间、能耗和管理开支。

SmartTier进行LUN级别的智能化数据存放管理。SmartTier统计和分析数据的活跃度，将不同活跃度的数据和不同特点的存储介质动态匹配。SmartTier通过数据迁移将活跃度高的“繁忙”数据迁移至具有更高性能的存储介质（如SSD硬盘），将活跃度低的“空闲”数据迁移至具有更高容量且更低容量成本的存储介质（如NL SAS硬盘）。

SmartTier的统计、分析和迁移活动基于SmartTier的实现策略和数据的性能要求。在统计、分析、迁移活动期间，不会对现有业务连续性和数据可用性造成影响。

SmartQoS技术

- SmartQoS是一种性能特性，通过动态地分配存储系统的资源来满足某些应用程序的特定性能目标。

主机IO请求
对IO请求分类
进入等待队列等待
按照规定的速率向桶内放入令牌
令牌
令牌桶
到令牌桶获取令牌
获得调度处理

 Page 41 HUAWEI

随着存储技术的不断进步，存储系统能够提供的存储容量越来越大，越来越多的用户选择将不同的应用程序后端存储部署在同一台存储设备上，但同时也带来了如下问题：

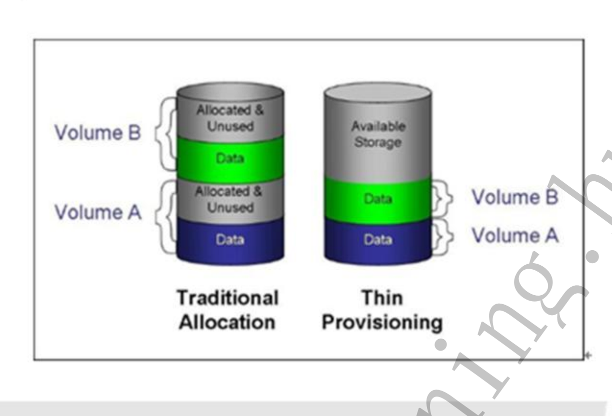
- 不同应用程序之间由于业务模型和I/O特征不同相互影响，导致存储系统整体性能受到影响。
- 不同应用程序相互争抢系统带宽和IOPS资源，关键业务性能无法得到保证。

SmartQoS技术能够帮助用户合理利用存储系统的资源，同时保证关键业务的性能。SmartQoS特性允许用户根据应用程序数据的一系列特征(IOPS或占用带宽)对每一种应用程序设置特定的性能目标。存储系统根据设定的性能目标，动态分配存储系统的资源来满足特定应用程序的服务级别要求，优先保证关键性应用程序服务级别的需求。

SmartQoS技术基于令牌桶原理实现，用户每配置一个SmartQoS策略，系统会根据用户设置的性能目标生成一个令牌桶。按照用户配置的性能目标周期性向令牌桶中放入一定数量的令牌。每一个受这个SmartQoS 策略控制的I/O请求都必须从令牌桶中获得一个令牌才能得到处理，如果令牌桶中的令牌取空，则只能在等待队列中等待系统下一次放入令牌。

Smart Thin技术

- SmartThin能够实现按需分配存储空间。在存储空间配额范围内，应用服务器用到多少空间，存储系统才给它分配多少空间，从而节省了宝贵的存储资源。



 Page 42 

- SmartThin是一种按需分配的方式来管理存储空间的技术，利用虚拟化方法减少物理存储部署，可最大限度提升存储空间利用率。
- Smart Thin提供按需分配的存储管理方式。
- Smart Thin提供精简LUN，按用户的实际需求分配物理存储空间。
- Smart Tier、Smart Thin、Smart Qos可以同时启用。

操作分级管理

- Level 0：无影响，不进行处理
- Level 1：提示
- Level 2：警告
- Level 3：危险

Level 3:系统复位/系统重启/系统offline

Level 2:业务中断/数据丢失/数据不一致

Level 1:性能降低

Level 0:系统配置改变但无危害

系统

业务与数据

性能

无影响

 Page 43 HUAWEI

通过对产品所有软件操作带来的影响进行分析（FMEA），按影响级别进行对应的防误操作设计，有效预防由于人为因素引起的错误操作。

根据系统、业务与数据、性能三个关键因素对操作进行分级，有效区分操作对存储阵列的影响，根据不同级别的操作给出对应的提示信息，确保不会出现误操作。

注：FMEA：Failure Mode and Effects Analysis 失效模式与影响分析



目录



ISM简介

- ISM是一种集成存储管理平台，ISM基于SMI-S协议，可以管理多套设备。

OceanStor ISM

欢迎使用OceanStor ISM集成存储管理系统，进入管理系统前建议您使用“发现设备”功能，查找当前网络中的存储设备。

发现设备

快速学习OceanStor ISM操作方法
http://www.huawei.com

基于SMI-S标准的管理系统
OceanStor ISM集成存储管理系统可管理支持SMI-S标准的存储设备。

友好的操作界面
操作界面易用、简捷，使管理和维护工作更加轻松。

实时的信息采集
实时的显示、通知存储设备运行状态，方便了解设备的最新状态。

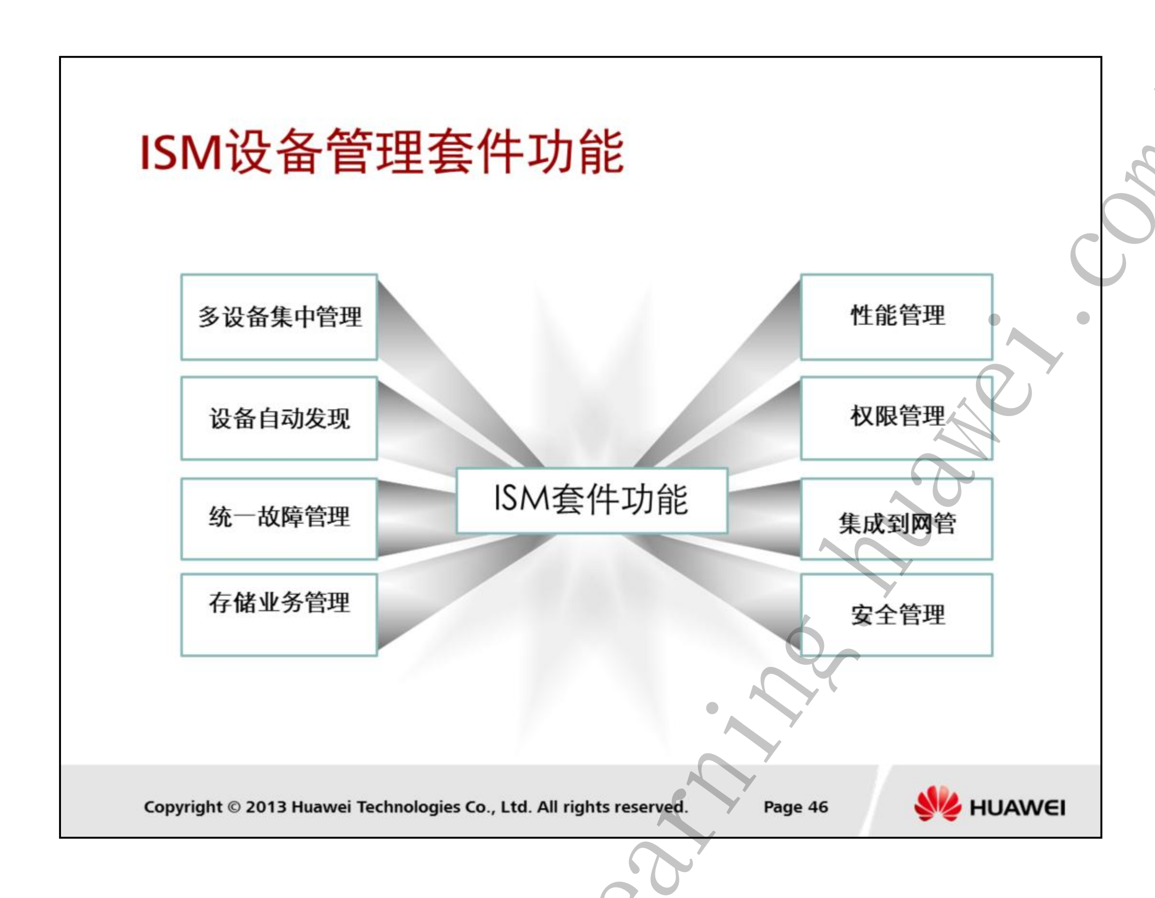
不再显示此对话框 关闭(L)

 Page 45



ISM（Integrated Storage Management）是华为科技有限公司针对存储系列产品开发的存储管理软件。目前包括ISM设备管理套件、ISM云存储管理软件、ISM存储网络管理软件。我们在配置阵列时一般只用到ISM设备管理套件。并且已经实现同时支持管理多种设备类型，目前包括N8000、VIS、存储阵列。

- ISM具备以下特性：
  - 友好的界面以及详细的系统提示信息；
  - 系统采用JWS（Java Web Start）方式部署，可方便的自动安装运行；
  - 提供用户鉴权，保证系统安全性。



- 多种设备集中管理：对华为的SAN、NAS、虚拟化等存储设备实现集中管理。
- 设备自动发现：自动发现网络中的存储设备，无须用户逐一添加设备。
- 统一的故障管理：网络中存储设备出现的任何故障都可以实时的通知到管理软件，并可以通过短信、Email发送给管理员。
- 存储业务管理：对存储空间分配、拷贝、快照、镜 像，提供向导功能指导管理员轻松完成业务配置。
- 性能管理：监控和显示系统的性能指标，识别系统的瓶颈，最多可监控一个月的性能数据。
- 权限管理：根据管理员级别，提供不同的操作权限。
- 集成到大型网管： 支持对主流大型网管系统的对接，可以与华为2000系列网管、Symantec CCS、HP OpenView、IBM TSM等网管进行集成。
- 安全管理：使用SSL，CHAP Secret、 MD5加密以保障系统的安全。

ISM安装

- 在管理终端的浏览器中输入所连接控制器的管理口的IP地址；
- ISM是基于JRE，因此维护终端要先安装JRE，才能安装ISM；



- ISM安装软件获取途径：
  - 官方网站；
  - 随机光盘；
  - 输入设备管理IP,通过设备上的下载连接下载安装。
- 管理终端必须与存储阵列控制器的管理网口实现IP互通。
- 目前ISM软件只支持windows的管理终端；
- 管理口的默认IP地址是A控 192.168.128.101； B控 192.168.128.102。

如果出现ISM不能安装或者登录的问题，建议卸载原有的JRE，安装ISM配套的或者版本更高的JRE。

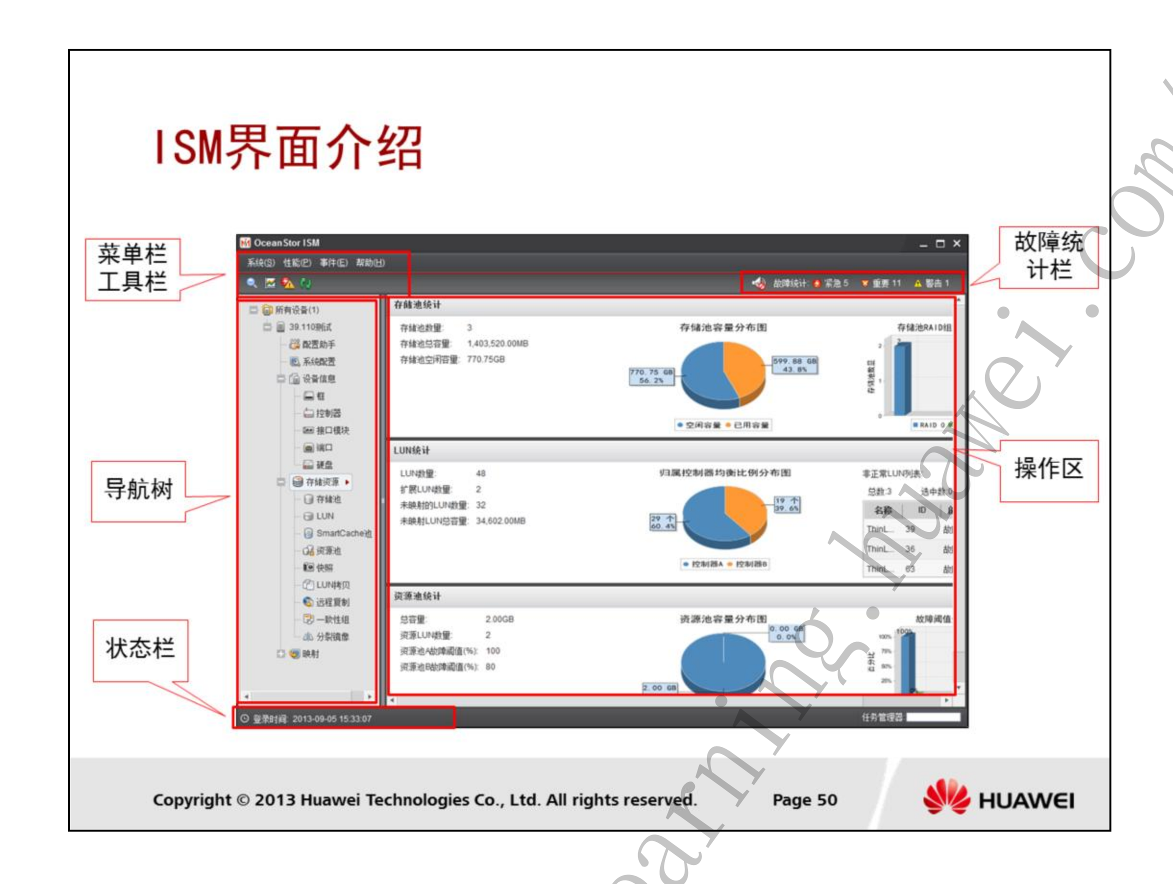


安装好ISM后，会在桌面出现ISM图标，双击ISM图标或者“开始”程序菜单中点击ISM图标。语言选择中，选择中文，点击确定。在弹出的欢迎界面中点击发现设备或者关闭欢迎界面，通过主界面发现设备。

勾选欢迎界面中的“不再显示此对话框”，下次启动的时候就不会再显示此页面。

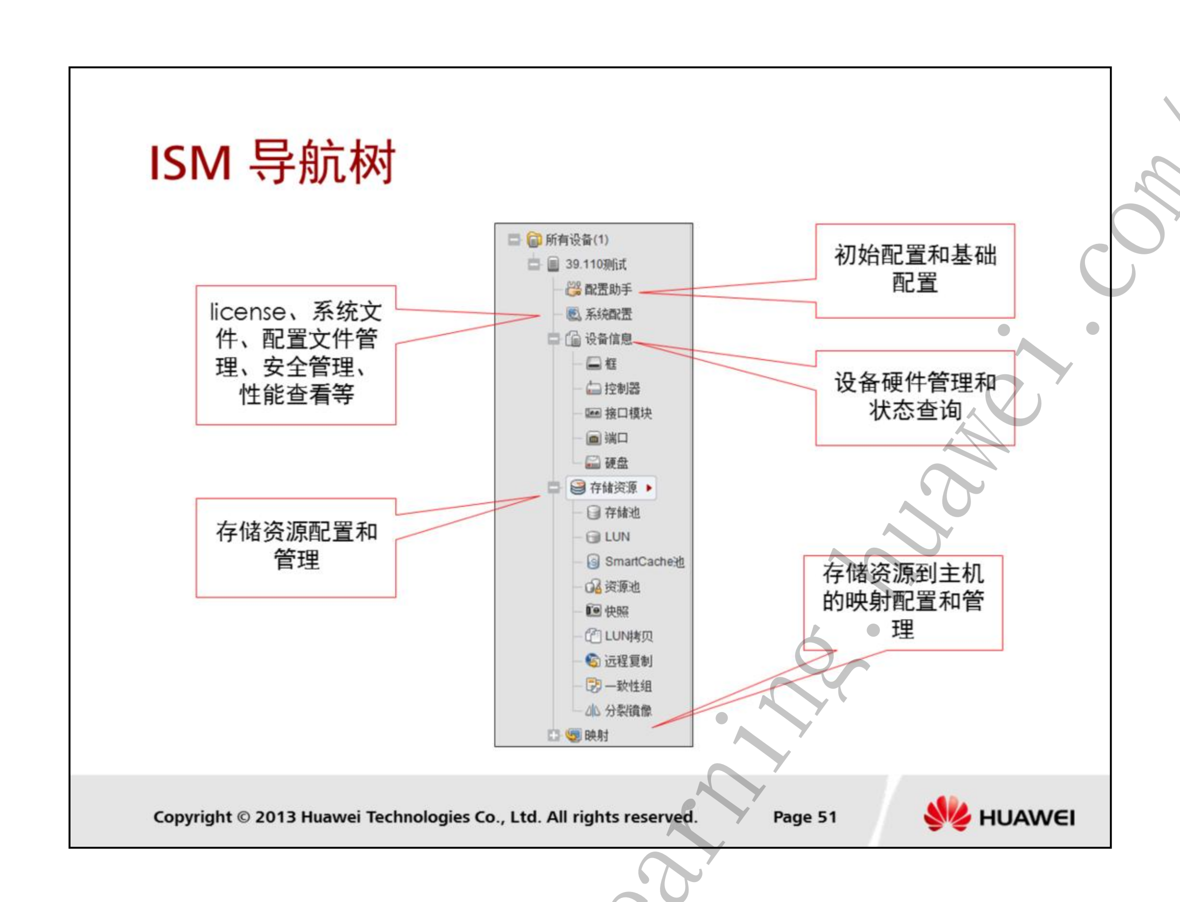


- 默认的控制器管理口IP地址
  - A控：192.168.128.101/24
  - B控：192.168.128.102/24
- 默认的用户名和密码
  - 用户名：admin
  - 密码：Admin@storage
- 发现新设备时发现方式有三种，分别为“指定IP地址” “指定IP地址段” “同一子网”；
- 管理网口可以更改，但是不能和业务网口在同一个网段



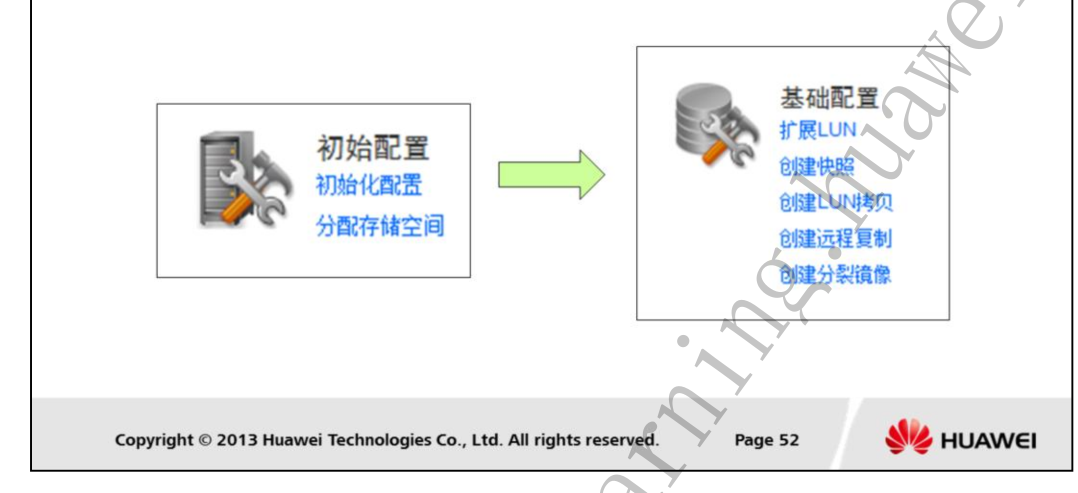
- ISM主界面主要内容包含：

1. 菜单栏：包括“系统”、“性能”、“事件”和“帮助” 4个菜单，主要功能如下：
   - 系统：提供发现设备、任务管理器、常用设置、刷新、锁定系统和退出ISM客户端等功能。
   - 性能：提供对存储系统的存储设备、前端主机端口、LUN和链路的性能统计功能。
   - 事件：提供对存储系统事件管理、客户端声音设置和客户端短消息通知设置功能。
2. 工具栏：提供了常用操作的快捷图标，主要包括 “发现设备” 、 “性能统计”、“事件管理”和 “刷新”。
3. 操作区显示相应导航树节点的信息，能够显示设备示意图和对应的设备信息或逻辑组件信息。设备图上可以单击对应的对象查看其属性。设备信息详细罗列了左侧节点对象的信息以及该节点下级的关键信息。包含：（1）设备基本信息（2）软件版本信息（3）主机/主机组统计信息（4）存储资源的统计信息 （5）硬件信息统计
4. 故障统计栏显示系统中的故障统计信息。通过故障统计栏可以查看系统当前存在的紧急、重要和警告三种级别故障信息的数量。 当存储设备有故障告警音提示时，通过故障统计栏左侧的 “关闭当前客户端声音”按钮，可以关闭ISM的事件提示音。
5. 状态栏显示任务管理和登录ISM时间信息。



ISM导航树显示了当前ISM所管理的存储系统的逻辑结构。通过选择不同的导航树节点，右侧操作区将显示设备示意图、对应的设备信息或逻辑组件信息。

通过导航树中相应模块的操作和配置可以完成对存储设备的业务配置和系统管理。



- 初始化配置向导

在ISM客户端发现设备后，可以通过初始化配置向导对存储设备进行相关设置，包括设备基本信息设置、设备时间设置、FC端口信息设置、iSCSI端口信息设置、事件通知设置、License管理。

- 分配存储空间向导

该向导将引导您创建存储空间，将存储空间分配给主机或主机组使用。

- 扩展LUN

当LUN的剩余容量不能满足业务需求，存储系统中存在空闲空间时，通过将空闲的空间合并到剩余容量不足的LUN上，对LUN的容量进行扩充。

- 创建快照

快照是源LUN在某个时间点的静态映像。快照被激活后，当源LUN的数据被修改时，修改前的数据会被拷贝到资源池中。如果对源LUN的数据进行了误操作，可以通过快照回滚操作将源LUN的数据恢复到快照时间点的状态，实现对源LUN数据的保护。通过该向导，您可以创建虚拟快照或定时快照。

- 创建LUN拷贝

LUN拷贝将源LUN的数据拷贝到目标LUN中。当源LUN中的数据出现故障时，可以使用目标LUN中的数据，或通过目标LUN恢复源LUN中的数据。通过该向导，您可以创建设备内或设备间的LUN拷贝。



- 创建远程复制

　　通过创建远程复制可以将主端存储设备上的数据以LUN为单位拷贝到远端存储设备上，实现数据的异地容灾。

- 创建分裂镜像

　　创建分裂镜像后，使用分裂镜像可以同时生成多份源LUN在分裂时间点的物理拷贝。对拷贝的数据可以进行多次独立的操作，不会对源LUN中的数据产生任何影响。



- 基本配置

实现对设备时间管理。通过License管理功能，可以浏览License信息、导入并激活License文件。

- 事件日志

实现对设备事件、告警等信息的提醒、保存等管理。

- 导入导出

实现对配置文件的导入和导出操作，对运行数据及系统志的收集和导出 操作。

- 性能监控

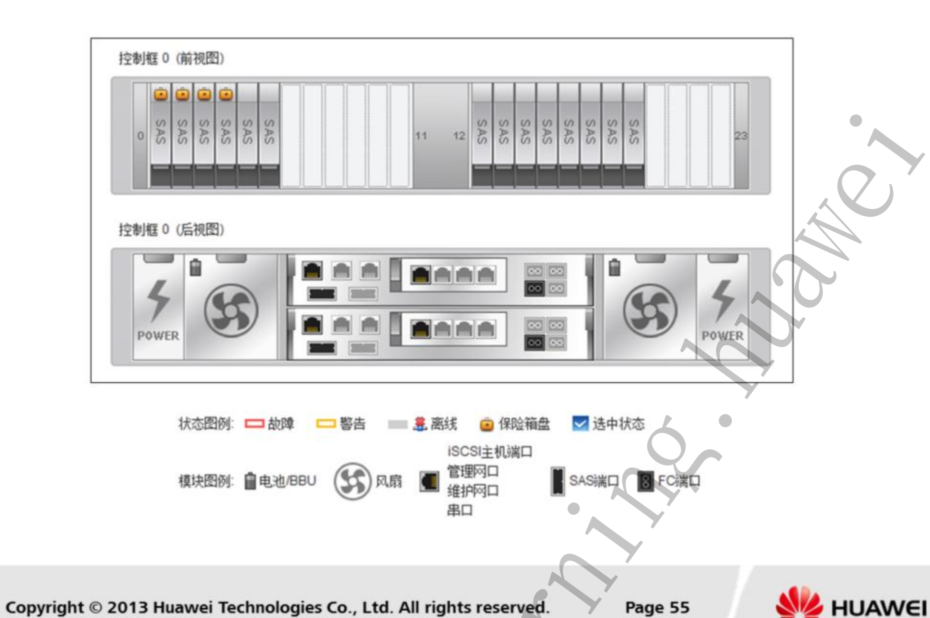
实现对设备性能监控的配置和监控信息的收集和文件管理。

- 安全

实现对用户管理、域认证管理和基于IP访问的管理。

- 高级

实现iSNS服务配置、修改iSCSI设备名称、软件升级、补丁管理等

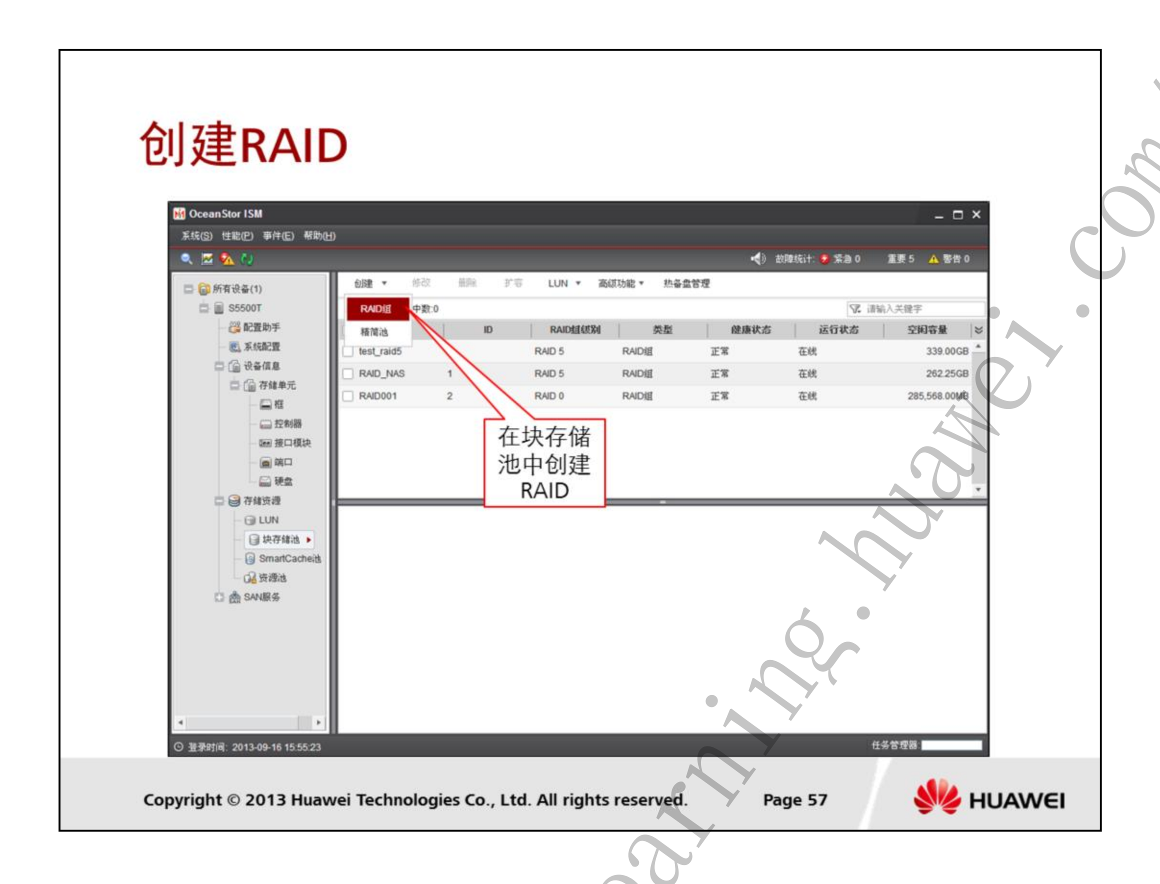


- 点击导航树“设备信息”，进入设备信息界面，主要是实时显示设备的各个组件状态，包括框、控制器、端口和硬盘等。点击设备图上的图标，可以显示该组件的详细信息。
  - 框：通过ISM的框管理功能，可以修改控制框或硬盘框的ID，查看控制框或硬盘框的信息以及框上各组件的信息。
  - 控制器：通过控制器管理功能，可以设置控制器的Cache高低水位来控制Cache中的脏数据的容量，修改控制器上的管理网口IP地址。
  - 接口模块：通过接口模块管理功能，可以对每个接口模块进行上电和下电操作。
  - 端口：FC、iSCSI端口的绑定、修改等操作，以及端口列表详细信息；
  - 硬盘：硬通过对硬盘的管理，可以为创建的RAID组设置热备盘，对硬盘实现定位功能，对硬盘进行扫描检查硬盘上数据的一致性。



目录





- 存储设备支持RAID 0、RAID 1、RAID 3、RAID 5、RAID 6、RAID 10和RAID 50，可以根据不同的应用创建不同级别的RAID组，各种RAID组适用的应用场景如下所示：
  - RAID 0：适用于以读写性能为第一要求，且数据保护要求最小的应用。
  - RAID 1：适用于以数据安全为第一要求的应用。
  - RAID 3：适用于对存储性能、数据安全和存储成本兼顾的应用。RAID组中的一个硬盘作为校验盘。任意一个成员盘失效，RAID组降级仍能正常工作。
  - RAID 5：适用于对存储性能、数据安全和存储成本兼顾的应用。RAID组中存在分散在不同条带上的奇偶校验数据。任意一个成员盘失效，RAID组降级仍能正常工作。
  - RAID 6：适用于对数据安全要求较高的应用。任意两个成员盘失效，RAID组降级仍能正常工作。
  - RAID 10：适用于既有大量数据需要存取，同时又对数据安全性要求严格的应用，如银行、金融、商业超市、仓储库房、各种档案管理等。
  - RAID 50：适用于既有大量数据需要存取，同时对数据安全性要求较高的应用。RAID 50是RAID 5与RAID 0两种技术的结合，至少需要6个空闲盘。
  - 创建RAID组时，RAID组成员盘的类型应保持一致。 如果存储系统中存在SATA硬盘和NL SAS硬盘时，可以选择SATA硬盘和NL SAS硬盘作为同一个RAID组的成员盘。

创建热备盘

勾选空闲盘后要点击箭头移动到热盘中

 Page 58 HUAWEI

- 热备盘管理
  - 热备盘是被指定用于替代RAID组内故障成员盘的硬盘，完成的任务是承载故障硬盘中的数据。通过热备盘管理可以将空闲盘设置为热备盘或将热备盘设置为空闲盘。
- 设置热备盘
  - 在导航树上，展开存储设备下的“存储资源”节点。
  - 单击“块存储池”节点。
  - 在右侧信息展示区选择“热备盘管理”。 系统弹出“热备盘管理”对话框。
  - 在“空闲盘”区域框中，勾选需要设置为热备盘的空闲盘。
  - 单击图标，空闲盘移动到“热备盘”区域框中。
  - 单击“确定”。 系统弹出“执行结果”对话框，提示操作成功。
  - 单击“关闭”。
- 取消热备盘
  - 在导航树上，展开存储设备下的“存储资源”节点。
  - 单击“块存储池”节点。
  - 在右侧信息展示区选择“热备盘管理”。 系统弹出“热备盘管理”对话框。
  - 在“热备盘”区域框中，勾选需要取消的热备盘。

创建LUN

创建LUN的时候要考虑好分条深度的大小，创建好后不能再修改

Page 59
HUAWEI

- 名称
  - 名称不能重复，只能包含半角的字母、半角的数字、“.”、“_”、“-”和简体中文字符，且“-”不能作为首字符，长度为1～32个字符（1个中文字符占3个字符长度）。
- 分条深度
  - 新建LUN的分条深度。分条深度是指在使用分条数据映射的硬盘设备中，条带内的块大小。也指在硬盘设备的单个成员盘区中，连续编址的虚拟硬盘块映射到连续编址的块的大小。分条深度取值包括“4 KB”、“8 KB”、“16 KB”、“32 KB”、“64 KB”、“128 KB”、“256 KB”和“512 KB”8种。
  - 由于分条深度影响到I/O性能，因此在不同的应用场景下应适当选择分条深度大小。例如：当系统应用于存储顺序数据较多的情况下，如存储媒体数据，建议设置较大的分条深度。推荐设置为64 KB。当系统应用于存储随机数据较多的情况下，如存储事务处理数据，建议设置较小的分条深度。推荐设置为32 KB。
- 归属控制器
  - 为了实现控制器A和控制器B的负载均衡，建议将存储系统中的LUN分配给不同的控制器。 说明： 当LUN所属RAID组的成员盘为SAS盘或FC盘时，LUN的归属控制器可以选择为“控制器A”、“控制器B”或“自动选择”。当LUN所属RAID组的成员盘为其他类型时，LUN的归属控制器只可以选择“控制器A”或“控制器B”。



总结

- 存储阵列系统的组成
- 华为存储阵列技术
- 华为存储阵列基础配置





思考题

- 存储系统由哪几部分组成？
- 存储阵列一般分为哪两部分？存储控制器的主要功能是什么？
- 什么是磁盘保险箱技术和磁盘预拷贝技术？主要解决什么问题？
- 什么是块虚拟化技术？与传统RAID技术有什么区别？
- 快照技术、LUN拷贝、远程复制技术有什么区别？
- 什么是Smart Tier技术、Smart Qos技术、Smart Thin技术？分别应用在什么场景？





习题

- 判断题
  1. 存储阵列控制框主要由控制器模块、风扇-BBU模块、电源模块以及接口模块等组成，是存储系统的核心部件。 （T or F）
  2. 存储阵列的内置BBU电池可保证在系统意外断电时，对Cache和硬盘框同时供电，让Cache中的数据写到硬盘中，实现Cache数据永久保存。 （T or F）
- 多选题
  1. 快照技术的特点主要包括以下哪几方面？（ ）

     A.通过快照可以实现完整的物理上的数据拷贝。

     B.快照可以灵活生成多个时间点的数据副本，在需要时可以快速地恢复数据。

     C.快照可以瞬间生成，不影响主机业务。

     D.快照可以根据用户自定义的备份策略来实现。

 Page 62



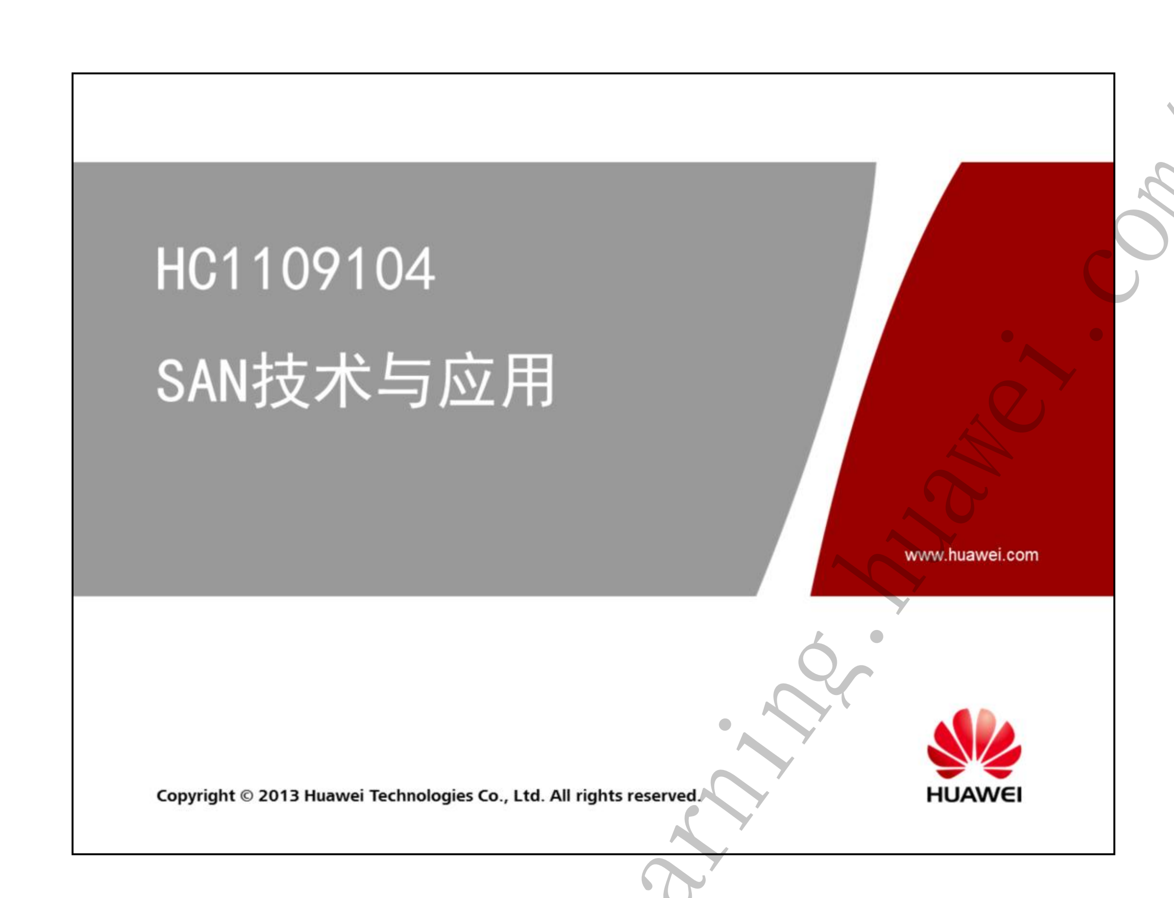
- 习题答案：
  - 判断题：1.T 2.F
  - 多选题：1.BCD

Thank you

www.huawei.com

HC1109104 SAN 技术及应用




HC1109104
SAN技术与应用
www.huawei.com
Copyright © 2013 Huawei Technologies Co., Ltd. All rights reserved.
HUAWEI



目标

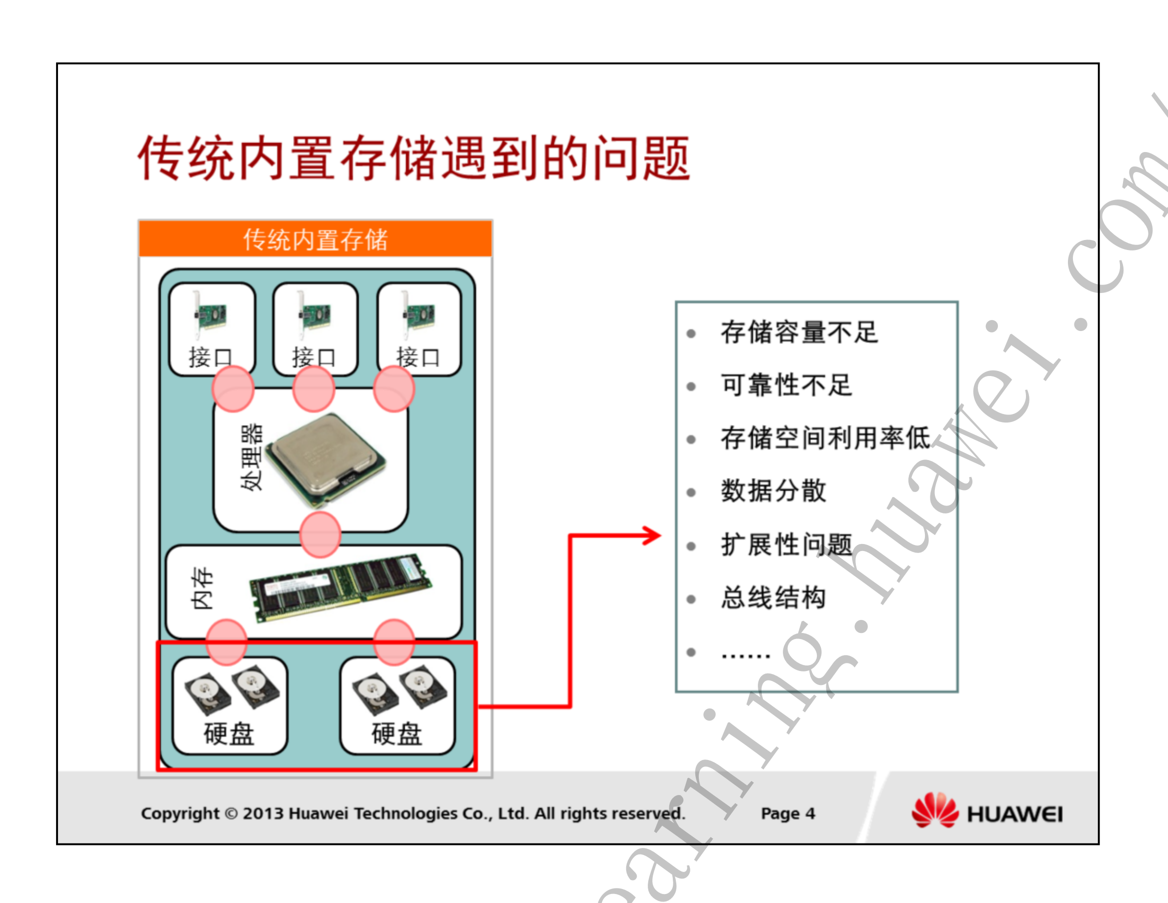
- 学习完本章节后，您将能够：
  - 掌握DAS存储基础知识
  - 理解SCSI协议
  - 掌握SAN存储的基础知识
  - 理解FC连接和协议封装
  - 掌握华为SAN存储的应用





目录





在传统的计算机存储系统中，存储工作通常是由计算机内置的硬盘来完成，而采用这样的设计方式，硬盘本身的缺陷很容易成为整个系统的性能瓶颈，并且，由于机箱内有限的空间，限制了硬盘数量的扩展，并且同时也对机箱内的散热、供电等提出了严峻的挑战。再加上不同的计算机各自为战，使用各自内置的硬盘，导致从总体看来存储空间的利用率较低，并且分散保存的数据也不利于数据的共享和备份工作。在传统的C/S架构中，无论使用的是何种协议，存储设备都直接与服务器相连接。在这样的结构下，对存储设备上所保存的所有数据的任何读写操作，都必须由服务器来进行，这样的处理方式给服务器带来了沉重的负担。外部存储系统的出现，彻底将服务器从繁琐的I/O操作中解放出来，使服务器更加专门化，使之仅仅承担应用数据的操作任务，以更充分的释放自身潜能。把存储设备从服务器中分离出来，使他们变成直接与网络连接的网络存储设备。所以存储区域网中，存储设备不再属于哪个应用服务器，从而可以对存储设备实施集中管理，使用户可以方便地共享存储资源。为网络上的应用系统提供丰富的存储资源，和快速、简便的访问方式。

存储网络建立了独立的基于网络的存储架构，增强了现有的计算拓扑架构。存储网络允许存储设备直接连接到现有网络上，也可以通过专门的存储网络进行连接，这一技术给传统的存储配置方案带来了两个重要的变化：

- 存储网络在存储设备、服务器以及客户机之间建立了更多的直接访问路径，从而使用户事务能够绕过大量的服务器I/O操作而直接与数据发生联系，从而避免了对服务器进行不必要的访问。
- 存储网络使得商务应用系统能够以更高的效率访问数据。换言之，存储网络使得应用系统能够更方便地共享数据，并赋予服务器更为强大的数据连接能力。

外置存储网络形态

- 存储网络几种常见类型

DAS | NAS | SAN

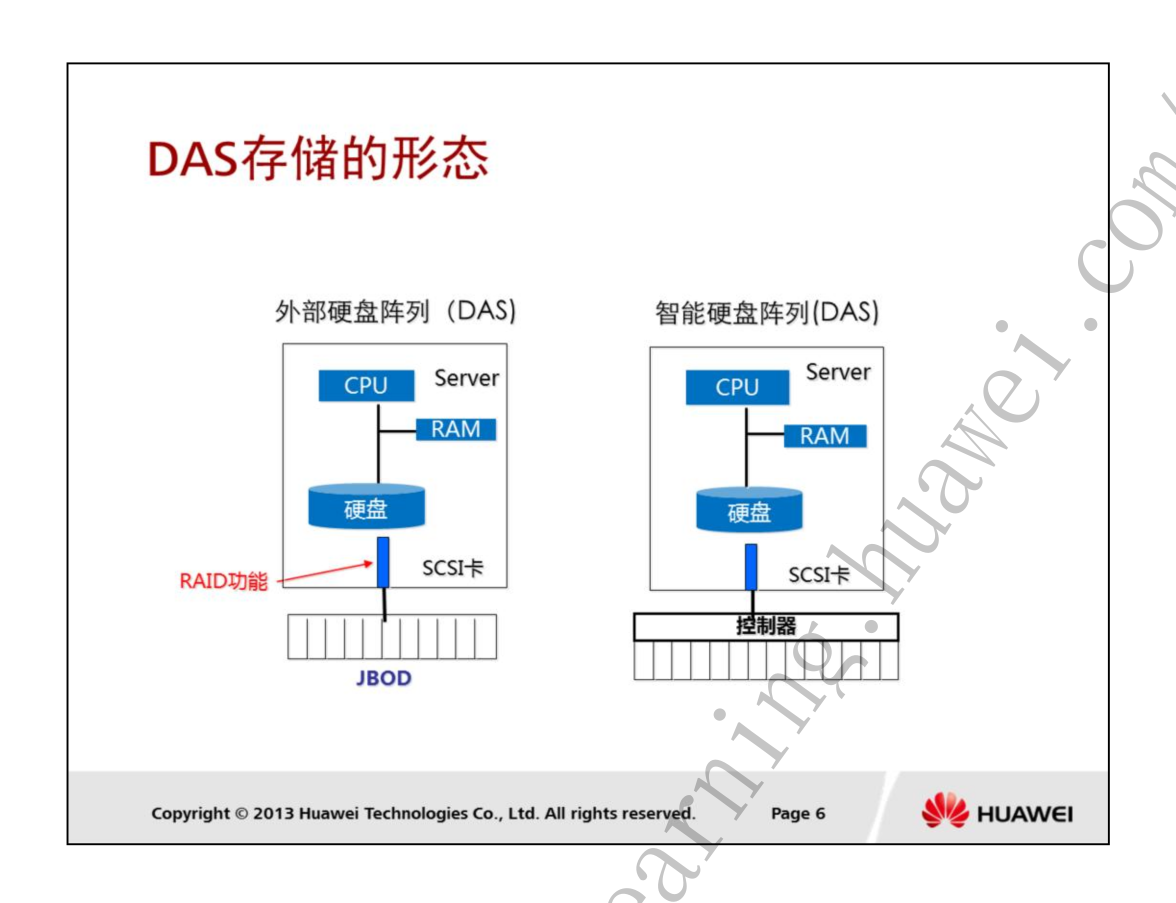
Application Software | File System | Storage | Network | FC/GbE

Page 5
HUAWEI

直接连接存储(Direct Attached Storage，DAS)：由于早期的网路十分简单，所以直接连接存储得到发展。随着计算能力、内存、存储密度和网络带宽的进一步增长，越来越多的数据被存储在个人计算机和工作站中。分布式的计算和存储的增长对存储技术提出了更高的要求。由于使用DAS，存储设备与主机的操作系统紧密相连，其典型的管理结构是基于SCSI的并行总线式结构。存储共享是受限的，原因是存储是直接依附在服务器上的。从另一方面看，系统也因此背上了沉重的负担，因为CPU必须同时完成磁盘存取和应用运行的双重任务，所以不利于CPU的指令周期的优化。

网络连接存储(Network Attached Storage，NAS)：局域网在技术上得以广泛实施，在多个文件服务器之间实现了互联，为实现文件共享而建立一个统一的框架。随着计算机的应用广泛，大量的不兼容性导致数据的获取日趋复杂。因此采用广泛使用的局域网加工作站族的方法就对文件共享，互操作性和节约成本有很大的意义。NAS包括一个特殊的文件服务器和存储设备。NAS服务器上采用优化的文件系统，并且安装有预配置的存储设备。由于NAS是连接在局域网上的，所以客户端可以通过NAS系统，与存储设备交互数据。另外，NAS直接运行文件系统协议，诸如NFS，CIFS等。客户端系统可以通过磁盘映射和数据源建立虚拟连接。

存储区域网络(Storage Area Networks，SAN)：一个存储网络是一个用在服务器和存储资源之间的、专用的、高性能的网络体系。它为了实现大量原始数据的传输而进行了专门的优化。因此，可以把SAN看成是对SCSI协议在长距离应用上的扩展。SAN使用的典型协议组是SCSI和Fiber Channel 。 Fiber Channel特别适合这项应用，原因在于一方面它可以传输大块数据，另一方面它能够实现远距离传输。SAN的市场主要集中在高端的，企业级的存储应用上。这些应用对于性能，冗余度和数据的可获得性都有很高的要求。

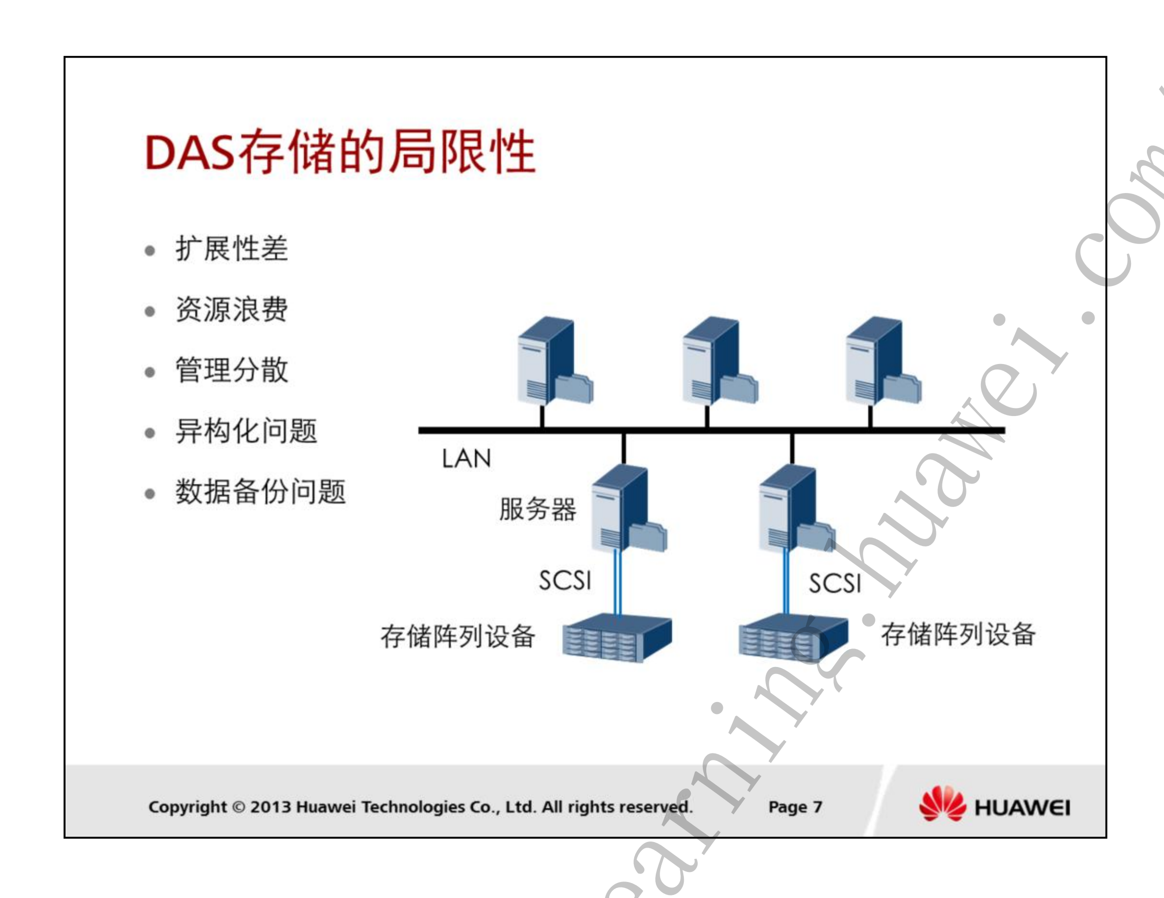


- 外部硬盘阵列：

　　JBOD 即Just a Bunch Of Disks，在逻辑上把几个物理磁盘串联在一起，其目的纯粹是为了增加磁盘的容量，并不提供数据安全保障。能够解决内置存储有限硬盘槽位，容量扩展不足的问题。但仍然是基于单硬盘存放数据，可靠性差。

- 智能硬盘阵列：

　　控制器中包含RAID功能、大容量Cache，同时使得磁盘阵列具有多种实用的功能，配置专用管理软件进行配置管理。



DAS存储方式实现了机内存储到存储子系统的跨越，但其也存在很多局限性：

1、扩展性差：服务器与存储设备采用直接连接的方式，当需要新增应用服务器时，只能为新增的服务器单独配置存储设备，造成重复投资。

2、浪费资源：存储空间无法充分利用，存在浪费。不同的应用服务器面对的存储数据量是不一致的，同时业务发展的状况也决定这存储数据量的变化。因此，出现了部分应用对应的存储空间不够用，另一些却有大量的存储空间闲置。

3、管理分散：DAS方式数据依然是分散的，不同的应用各有一套存储设备。管理分散，无法集中。

4、异构化严重：DAS方式使得企业在不同阶段采购了不同型号不同厂商的存储设备，设备之间异构化现象严重，导致维护成本据高不下。

5、数据备问题：DAS方式与主机直接连接，在对重要的数据进行备份时，将会极大地占用网络的带宽。

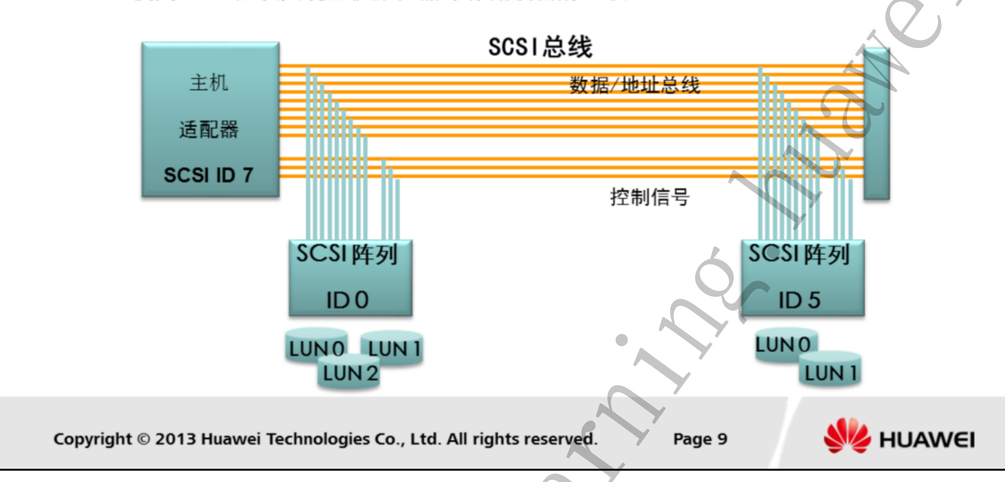
DAS特别适合于对存储容量要求不高、服务器的数量很少的中小型局域网，其主要的优点在于存储容量扩展的实施非常简单，投入的成本少而见效快。

DAS通常使用SCSI协议实现主机服务器与存储设备的互联。在下一小节将详细介绍SCSI协议。



目录

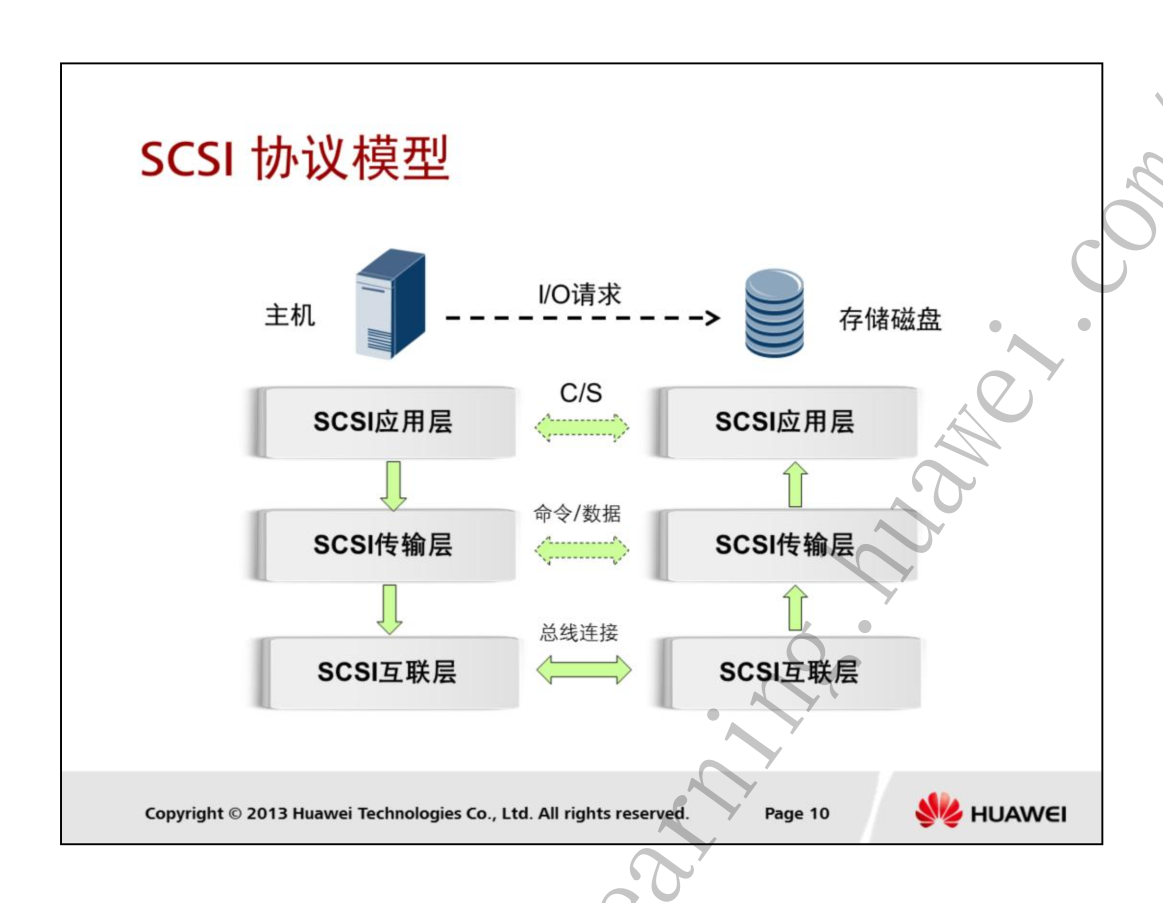




计算机与存储系统之间的通信是通过总线来完成的。总线就是从源设备传输数据到目标设备的路径。在最简单的情况下，控制器的高速缓存作为源，将数据传输给目标磁盘。控制器首先向总线处理器发出请求使用总线的信号。该请求被接受之后，控制器高速缓存就开始执行发送操作。在这个过程中，控制器占用了总线，总线上所连接的其它设备都不能使用总线。当然，由于总线具备中断功能，所以总线处理器可以随时中断这一传输过程并将总线控制权交给其它设备，以便执行更高优先级的操作。

计算机中布满了总线——从一个位置向另一个位置传输信息和电力的高速通道。例如，将MP3或数码相机连接到计算机时，您可能会使用通用串行总线 (USB)端口。对于存储图片、音乐等的小型电子设备，USB端口完全可以胜任传输数据和充电的工作。但是，这种总线还不足以同时支持整合计算机和服务器以及其他许多设备。

在这种情况下，就需要使用SCSI这样的总线。SCSI直译为小型计算机系统专用接口(Small Computer System Interface)是一种连结主机和外围设备的接口，支持包括磁盘驱动器、磁带机、光驱、扫描仪在内的多种设备。它由SCSI 控制器进行数据操作，SCSI控制器相当于一块小型CPU，有自己的命令集和缓存。SCSI是一种特殊的总线结构，可以对计算机中的多个设备进行动态分工操作，对于系统同时要求的多个任务可以灵活机动的适当分配，动态完成。

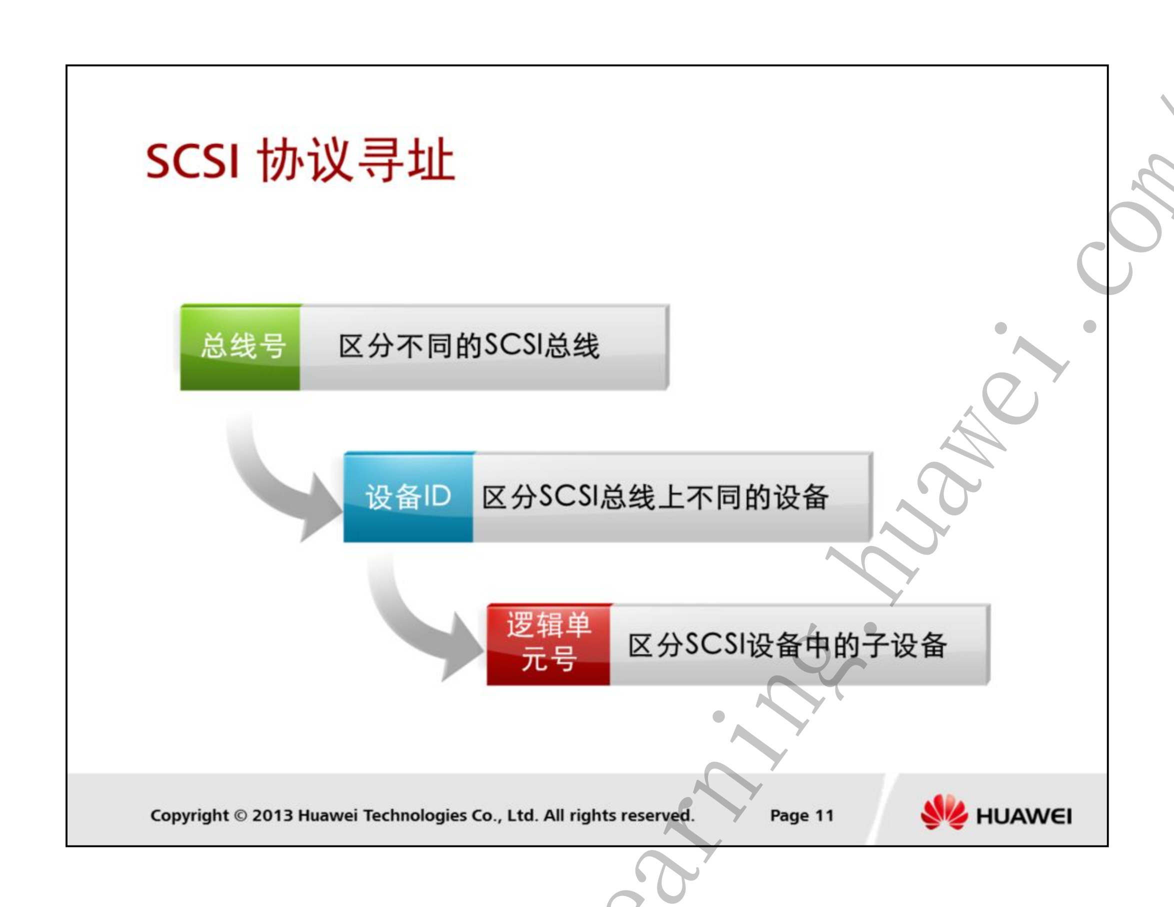


为了便于实现和理解SCSI的各个协议，SCSI 采取了分层结构。SCSI可分为三层，即SCSI应用层，SCSI传输层和SCSI互连层。

在应用层，SCSI协议采用C/S（客户/服务器）体系架构，SCSI协议客户端位于主机，代表上层应用程序、文件系统和操作系统发起I/O请求，SCSI设备服务器位于目标设备(如存储磁盘）中，对客户端I/O请求做出响应。 客户/服务器请求和响应通过其下层协议进行传输。

在传输协议层，SCSI设备之间通过一系列的命令实现数据块的传送，大致分成三个阶段：命令的执行，数据的传送和命令的确认。

SCSI互联层完成SCSI设备对总线的连接以及发送方和目标方的选择等功能。



为了对连接在SCSI总线上的设备寻址，SCSI协议引入了SCSI设备ID和逻辑单元号LUN。在SCSI 总线上的每个设备都必须有一个唯一的ID，其中包括服务器中的主机总线适配器也拥有设备ID。每条总线最多可允许有8个或者16个设备ID。

存储设备可能包括若干个子设备，如虚拟磁盘、磁带驱动器和介质更换器等。 因此SCSI引入了逻辑单元号，以便于对存储设备中的子设备进行寻址。

传统的SCSI适配卡连接单个总线，相应的只具有一个总线号。一个服务器可能配置了多个SCSI控制器，从而就可能有多条SCSI总线。在引入存储网络之后，每个光纤通道HBA（Host Bus Adapter）或iSCSI（Internet SCSI）网卡也都连接一条总线，必须对每一条总线分配一个总线号，在他们之间依靠不同的总线号加以区分。因此，我们可以使用一个三元描述标识一个SCSI目标：总线/目标设备/逻辑单元号。



右键点击我的电脑，选择管理，再选择磁盘管理，选择映射的磁盘，右键选择属性，在常规选项卡中查看位置，就是SCSI ID信息。

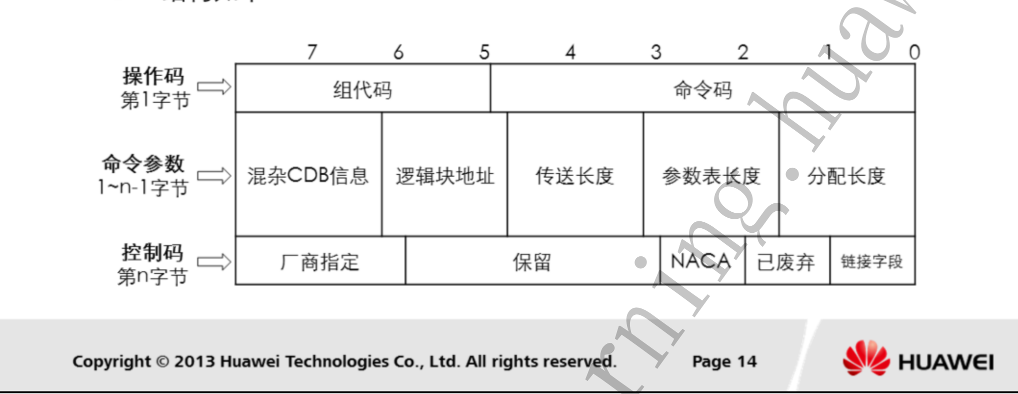
Linux系统查看SCSI设备ID方式

- 在Linux系统的命令行中，输入命令lsscsi：

```
linux-suse-icy:/proc/scsi # lsscsi
[0:0:0:0]    disk    ATA      ST3160318AS      CC65  /dev/sda
[0:0:1:0]    cd/dvd  HL-DT-ST DVD-ROM DH10N    0M10  /dev/sr0
[2:0:0:0]    disk    HUAWEI   S5500T           2105  /dev/sdb
```


Page 13
HUAWEI

每一行前面方括号中条目分别是SCSI host、channel、target number、LUN tuple，每个元素由冒号分开。当有多个SCSI设备条目时按元组升序排列。



在互连层完成SCSI设备对总线的连接，以及发送方和目标方选择的基础上，传输层协议执行实际的数据传输。传输协议的运行过程包括发送命令、传输数据和对命令执行的确认。发起方通过命令描述块（CDB）向目标方发送具体的命令。命令描述块有定长和不定长两种格式，而定长格式又有6、10、12、16字节不同的长度规定。

- 操作码

操作码是所有命令描述块都有的，它总是被放在命令描述块的开头一个字节。5-7位是组代码，指示该命令具体属于哪个命令组，它决定CDB的长度，0-4位则是具体的命令代码。8比特在理论上共有256个可能的操作码。

- 命令参数
  - 混杂CDB信息:该参数表示与具体的CDB相关的信息，如表示逻辑设备号。
  - 逻辑块地址: 该地址是逻辑单元(比如磁盘)中的起始操作块的位置。
  - 传送长度: 该长度表示命令所请求的传送量，通常是块数。
  - 参数表长度: 表示需要传送到存储设备的参数的长度，0表示不需要传递参数。
  - 分配长度: 分配长度表示应用客户为缓冲区分配的最大长度，根据具体的CDB类别，可能是字节数，也可能是块数。

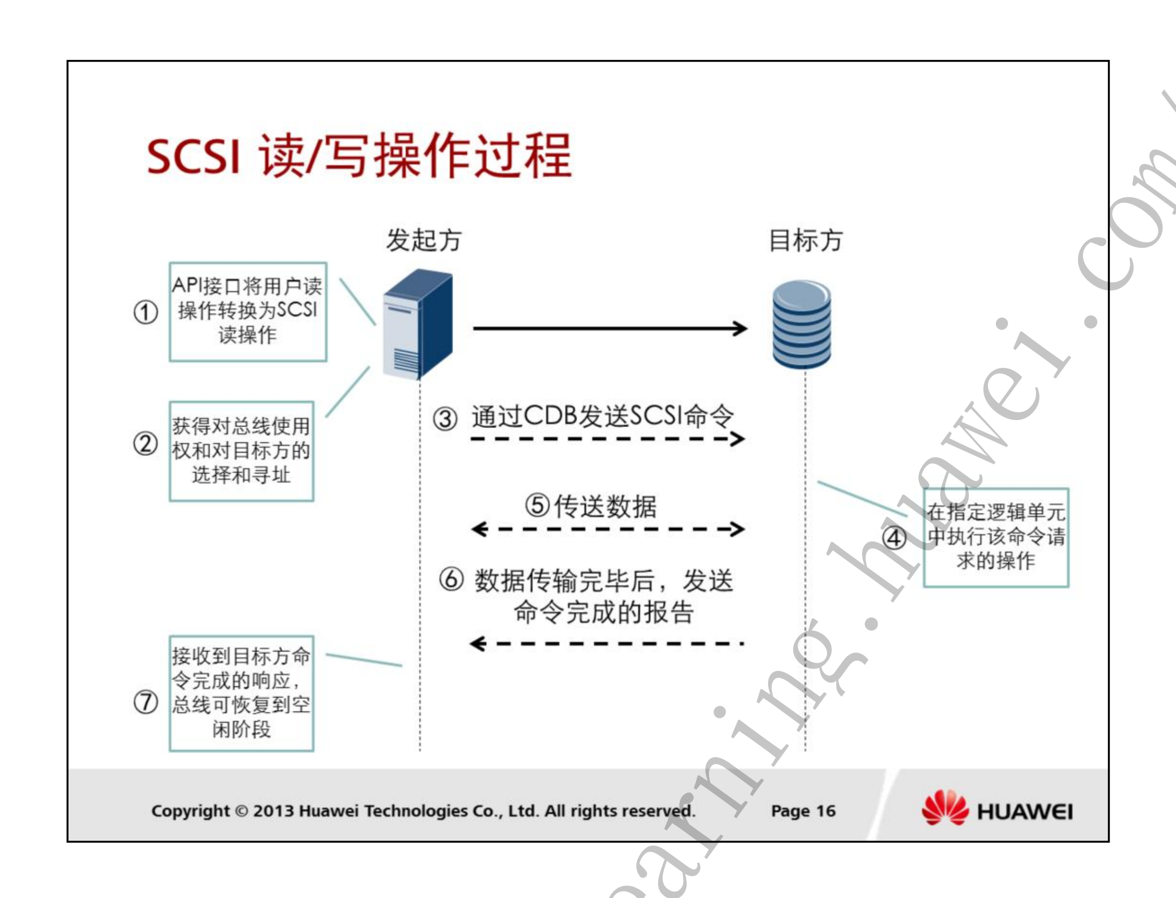
SCSI 命令描述块（CDB）

- 发起方通过命令描述块（command description block）向目标方发送具体的命令。
- SCSI基础命令规范SPC（SCSI Primary Commands，SCSI基础命令）定义了CDB的标准。
- CDB结构如下：



 Page 15 

- 控制码：所有CDB格式的最后一个字节。
  - NACA比特是为了让应用能够事先声明哪些命令执行的错误或异常需要善后处理，指定当指令返回CHECK CONDITION状态的时候，自动应急处理（auto contingent allegiance，ACA）是否会被创建。
  - 链接比特可以被用作跨越多个指令延续任务。比特为1时表示发起方请求跨越多个SCSI指令延续任务。



主机需要从存储设备获取数据，SCSI读操作需要完成以下步骤：

1. 主机操作系统需首先将用户的读取操作通过SCSI I/O的应用程序编程接口(Application Programming Interface，API)转化为SCSI的读操作，并在操作完成后通过相应的API返回响应的值。
2. 发起方SCSI总线由空闲阶段进入总线仲裁和选择阶段，获得对总线使用权以及对目标方的选择和寻址。
3. 发起方通过CDB向目标方发送SCSI的读命令。
4. 目标方接收到该命令，通过设备管理器在指定的逻辑单元中执行该命令请求的操作。
5. 目标方以字节为单位向发起方传送所需要的数据。
6. 在数据传输完毕后，目标方向发起方发送命令完成的报告。
7. 发起方接收到命令完成的响应，总线可恢复到空闲阶段。

SCSI的写操作过程与读操作过程类似，但数据传送的方向不同，它把数据从发送方向目标方传送。

SCSI协议的常见类型

接口模式	传输率（MB/s）	数据频宽（bits）	可连接设备数
SCSI-1	5	8	8
SCSI-2	10	8	8
SCSI-3(Ultra SCSI)	20	8	8
SCSI-3(Ultra Wide SCSI)	40	16	16
Ultra 2 SCSI	80	16	16
Ultra-160 SCSI	160	16	16
Ultra-320 SCSI	320	16	16
Ultra-640 SCSI	640	16	16

 Page 17

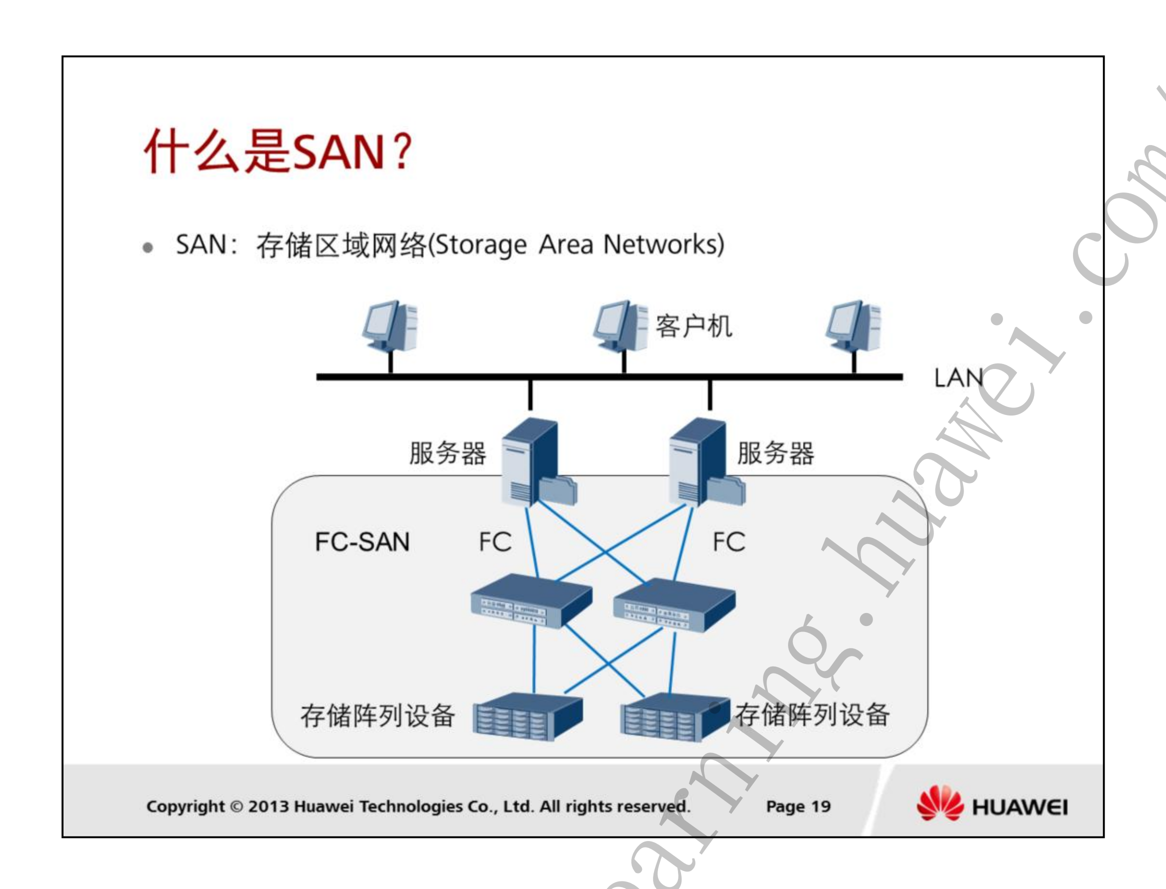


- 目前常见的SCSI类型及各自性能特征如下：
  - SCSI-1：它是最早的SCSI接口，它的特点是支持同步和异步SCSI外围设备，支持7台8位的外围设备，最大数据传输率为 5MB/s。
  - SCSI-2：它是SCSI-1的后续接口，是1992年提出，也称为 Fast SCSI。如果采用原来的8位并行数据传输则称为“Fast SCSI”，它的数据传输率为10MB/s，最大支持连接设备数为7台。
  - SCSI-3：它是在SCSI-2之后推出的。如果采用原来的8位并行数据传输时称为“Ultra SCSI”，它的数据传输率为20MB/s，最大支持连接设备数为7台。在将并行数据传输的总线带宽提高到16位后出现了“Ultra Wide SCSI”，它的传输率又成倍提高，即达到了40MB/s，最大支持连接设备数为15台。
  - Ultra2 SCSI：它是在Ultra SCSI的基础上推出的SCSI接口类型。
  - Ultra160 SCSI：它是Ultra2 SCSI的更新接口，使用和Ultra2 SCSI 完全一样的接口电缆及终结器，但是由于 Ultra160 SCSI采用双缘传输频率（Double Transition Clocking），而Ultra2 SCSI采用的是单缘传输频率，因此Ultra160 SCSI 的传输率是前者的两倍，即160MB/s。
  - Ultra320 SCSI：它的技术规范为“SCSI-3 SPI-4”。Ultra320 SCSI 单通道的数据传输速率最大可达320MB/s。
  - Ultra640 SCSI：它的技术规范为“SCSI-3 SPI-5”。Ultra640 SCSI 的数据传输速率最大可达640MB/s。



目录





存储区域网络(Storage Area Networks，SAN)：一个存储网络是一个用在服务器和存储资源之间的、专用的、高性能的网络体系。SAN是独立于LAN的服务器后端存储专用网络。SAN采用可扩展的网络拓扑结构连接服务器和存储设备，每个存储设备不隶属于任何一台服务器，所有的存储设备都可以在全部的网络服务器之间作为对等资源共享。

SAN主要利用Fibre Channel protocol（光纤通道协议），通过FC交换机建立起与服务器和存储设备之间的直接连接，因此我们通常也称这种利用FC连接建立起来的SAN为FC-SAN。FC特别适合这项应用，原因在于一方面它可以传输大块数据，另一方面它能够实现较远距离传输。SAN主要应用在对于性能、冗余度和数据的可获得性都有很高的要求高端、企业级存储应用上。

随着存储技术的发展，目前基于TCP/IP协议的IP-SAN也得到很广泛的应用。IP-SAN具备很好的扩展性、灵活的互通性，并能够突破传输距离的限制，具有明显的成本优势和管理维护容易等特点。针对IP-SAN技术的详细内容将在下一章讲解，接下来重点介绍SAN技术。

SAN典型组网

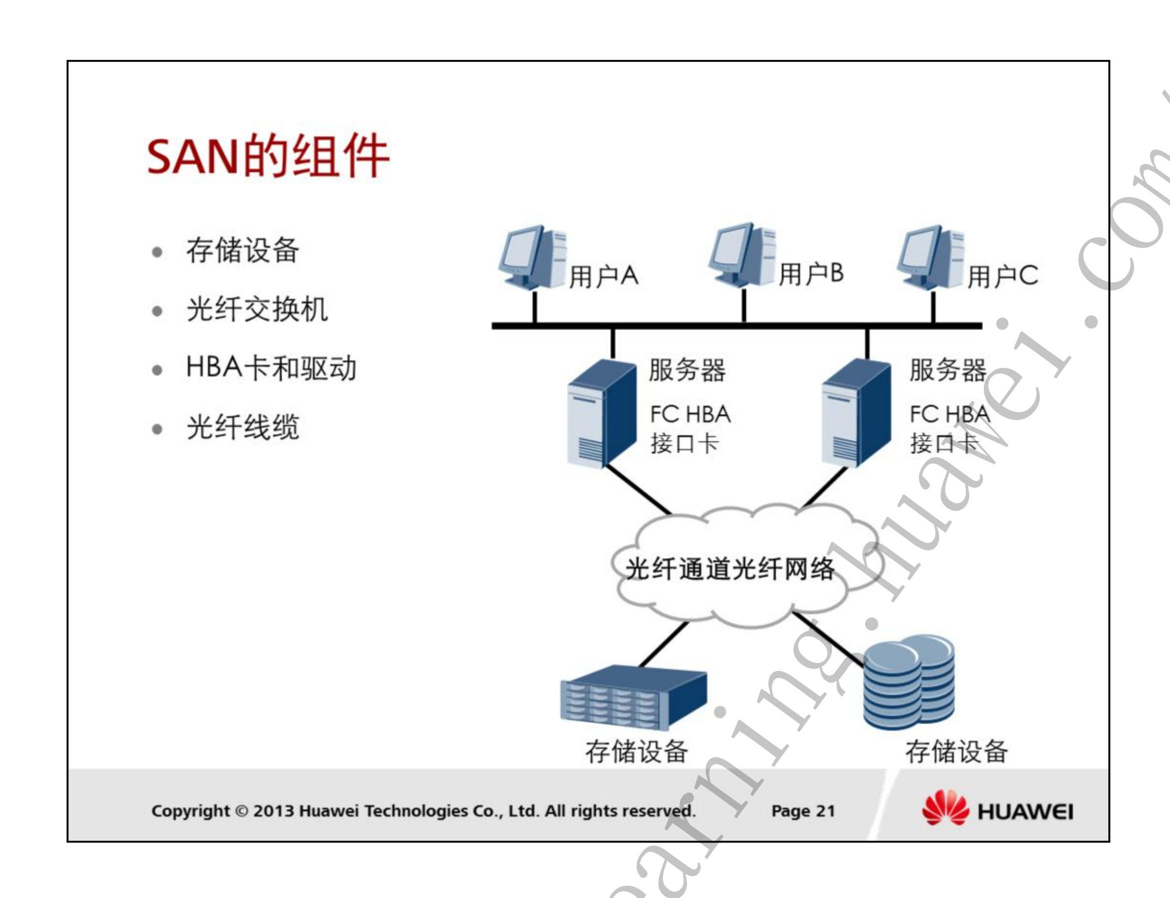
直连
应用服务器
光纤卡
FC HBA
存储设备

单交换
应用服务器 应用服务器
光纤交换机
存储设备

双交换
应用服务器 应用服务器
光纤交换机 光纤交换机
存储设备

 Page 20 HUAWEI

- 直连：主机与存储之间通过FC HBA卡连接，这种组网方式简单、经济，但较多的主机分享存储资源比较困难；
- 单交换：主机与存储之间由一台FC交换机连接，这种组网结构使多台主机能共同分享同一台存储设备，扩展性强，但交换机处存在单点故障；
- 双交换：同一台主机到存储阵列端可由多条路径连接，扩展性强，避免了在交换机处形成单点故障。

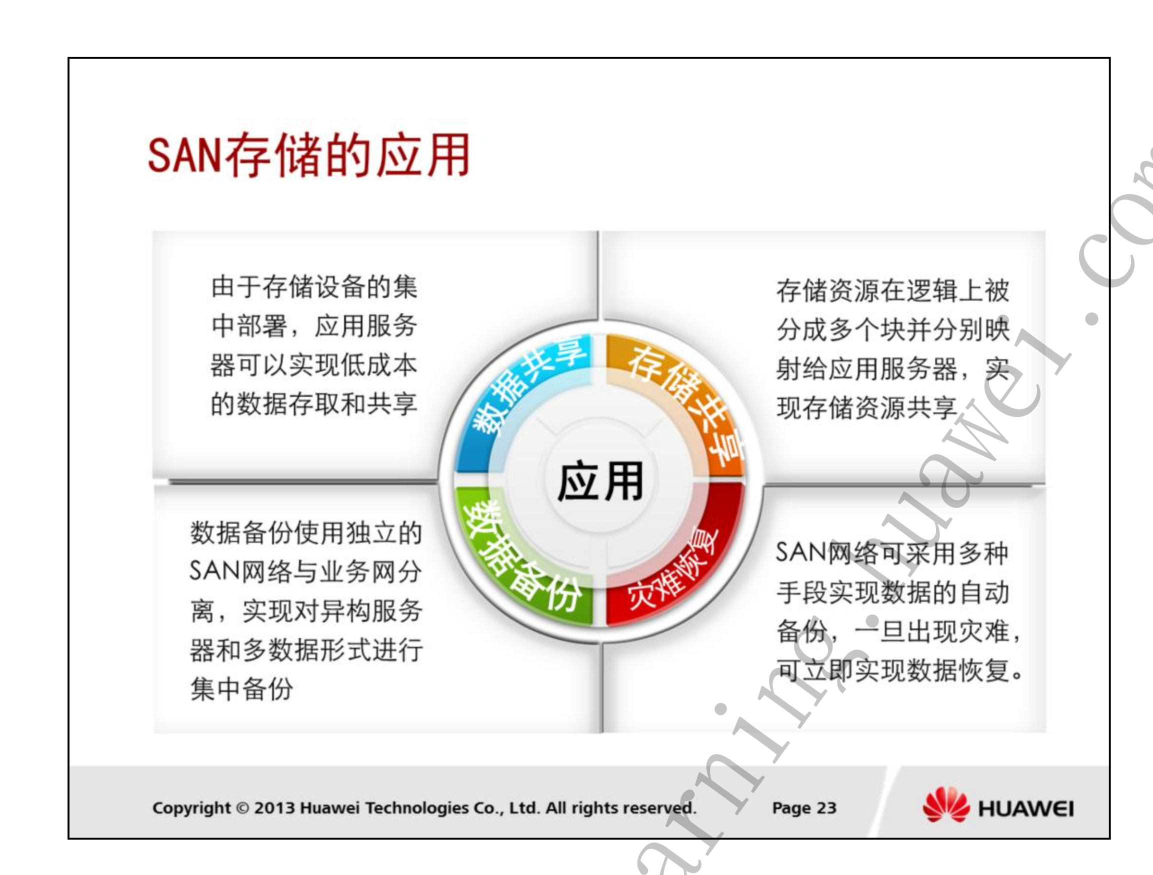


光纤存储区域网络由四个主要的部分组成，这些组件覆盖了I/O操作、存储系统以及所支持的工作负荷的各个主要方面。同时，在SAN技术中，还需要一些其它组件进行集成，以构建完整的解决方案。在考虑SAN的基础设施建设时，我们需要谨慎地考虑构成SAN基础设施的各个不同的组件，因为在SAN中，这些组件或者相互独立运作，或者相互依赖以协同工作。



与传统DAS存储相比， SAN存储网络具备非常明显的优势：

1. 基于FC光纤介质，拥有千兆位的存储带宽，更适合大容量数据块业务高速处理的要求，目前主流带宽为8G。
2. 由于SAN存储网络中存储设备（如磁盘阵列，磁带库等）集中部署，可以实现对设备的集中管理，同也可以通过远程方式登录设备实现远程管理，管理方式更加灵活。
3. 存储资源集中统一部署，分别映射给各应用服务器，实现存储资源的共享，同时可以根据各应用服务器对存储资源的需求为其动态的分配资源，实现存储资源的动态共享。
4. 在SAN存储网络，数据的传输、复制、迁移、备份等在SAN网内高速进行，不需占用业务网络WAN/LAN的带宽资源。
5. 在SAN存储网络环境下，扩展存储资源变得非常容易，只需要增加新的存储设备到SAN存储网络中即可，实现平滑的扩容。新增的存储资源可以直接映射给应用服务器使用。
6. 由于SAN存储网络使用的FC协议实现了对SCSI协议的封装，因此可以实现对以前的各种SCSI存储设备的兼容，在异构环境下，更能体现其优势。



SAN存储网络主要应用在：

- 对响应时间、可用性和可扩展性要求高的关键任务数据库应用。
- 对性能、数据完整性和可靠性要求高的集中存储备份，以保证关键数据的安全，可极大地提高企业数据备份和恢复操作可靠性和可扩展性。
- 海量存储的应用环境。例如，图书馆、银行、证券、中大型企业或组织的数据中心。
- 支持服务器及其连接设备之间提供光纤通道高性能和扩展的距离。

SAN与DAS的区别

项目	DAS	SAN
协议	SCSI协议	FC协议
应用	对存储容量要求不高、服务器数量很少的中小型局域网	关键数据库、集中存储、海量存储、备份、容灾等中高端存储应用环境
优势	部署简单、投资少	高可用性、高性能、高扩展性、兼容性、集中管理
劣势	可扩展性差、资源浪费、不易管理、性能瓶颈	投资相对较高

 Page 24

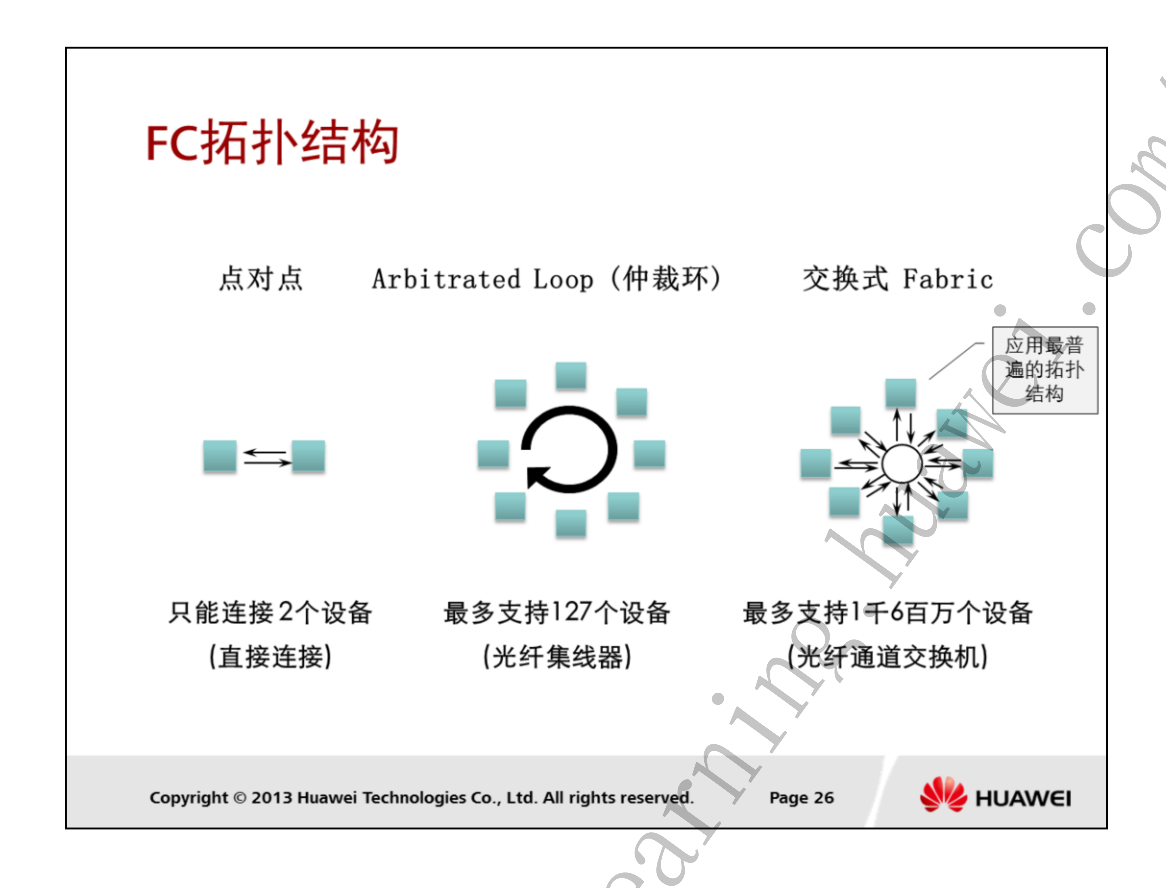


SAN存储网络目前在市场上已经得到广泛的应用。SAN存储网络的组网连接以及FC协议将在下一节进行详细介绍。



目录





FC主要有三种拓扑结构，用以描述各个节点的连接方式。光纤通道术语中的“节点”是指通过网络进行通信的任何实体，而不一定是一个硬件节点。这个节点通常是一个设备，比如说一个磁盘存储器，服务器上的一个主机总线适配器或者是一个光纤网交换机。

- 点到点式

两个设备背对背直接连接。这是最简单的一种拓扑，连接能力受限。

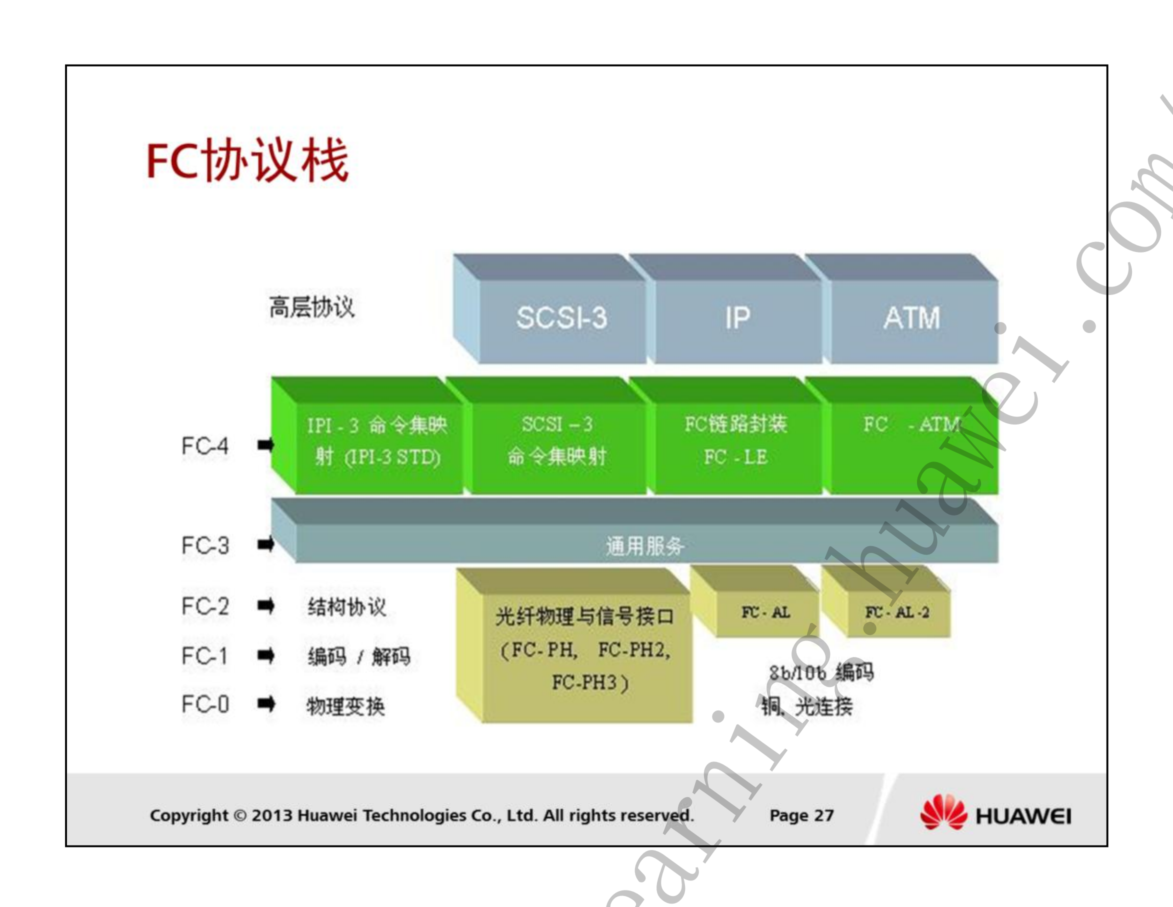
- 仲裁环式

这种设计方式中，所有设备连接在一个类似于令牌环的环路上。这个环路中添加或者移除一个设备会导致环路上所有活动中断。一个设备的故障导致整个环路不能进行工作。光纤通道集线器能够用于将众多设备连接到一起形成一个逻辑上的环路，并且能够旁路故障节点，使得环上节点的故障不会影响整个环路的通信。一个环路也可以通过使用线缆直接将节点一个接一个的连接成一个环而实现。最小的环路只包含两个节点，这种结构看起来和点到点式连接近似，它们的区别在很大程度上取决于各自的协议。

- 光纤交换式

所有的设备或者设备环都被连接到光纤网交换机上，与现有的以太网的实现形式在概念上是类似的。这种拓扑结构相对于点到点和仲裁环的优势在于：

- 交换机对结构形式进行管理，提供了最好的互联形式。
- 多对节点可以同时通信。
- 各个节点的故障是孤立的，不会危及其他节点的工作。



FC开发于1988年，最早是用来提高硬盘协议的传输带宽，侧重于数据的快速、高效、可靠传输。到上世纪90年代末，FC-SAN开始得到大规模的广泛应用。光纤通道的主要部分实际上是FC-2。其中从FC-0到FC-2被称为FC-PH，也就是“物理层”。光纤通道主要通过FC-2来进行传输，因此，光纤通道也常被成为“二层协议”或者“类以太网协议”。

光纤通道的数据单元叫做帧。即使光纤通道本身就有几个层，大部分光纤通道是指第2层协议。一个光纤通道帧最大是2148字节，而且光纤通道帧的头部比起广域网的IP和TCP来说有些奇怪。光线通道只使用一个帧格式来在多个层上完成各种任务。帧的功能决定其格式。相比我们在IP世界中的概念，光纤通道帧格式是奇特而且奇妙的。

光纤通道帧起始于帧开始（SOF）标志，随后是帧头部，这个一会进行描述。数据，或光纤通道内容，紧随其后，然后是帧结束（EOF）。这样封装的目的是让光纤通道可以在需要时被其他类似于TCP的协议所承载。

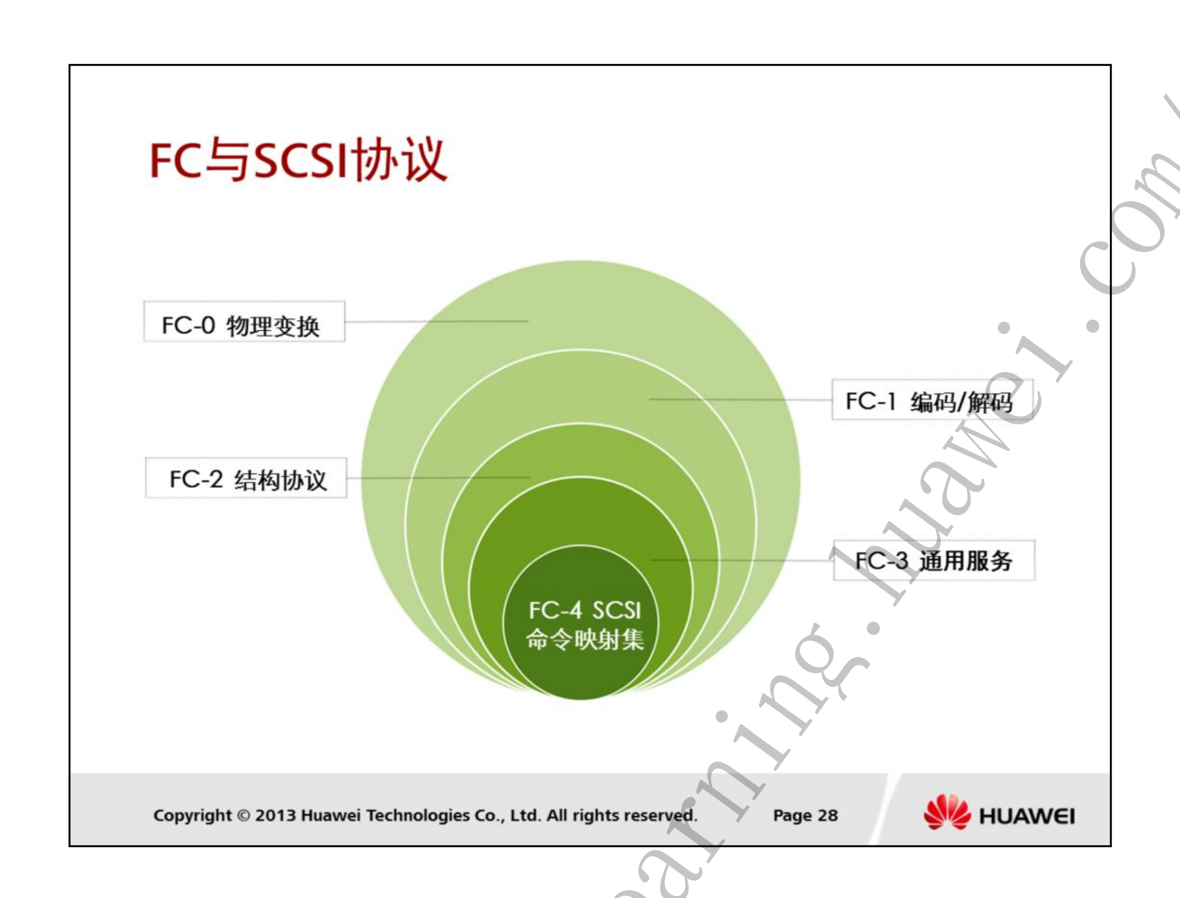
FC-0：物理层，定制了不同介质，传输距离，信号机制标准，也定义了光纤和铜线接口以及电缆指标；

FC-1：定义编码和解码的标准；

FC-2：定义了帧、流控制、和服务质量等；

FC-3：定义了常用服务，如数据加密和压缩；

FC-4：协议映射层，定义了光纤通道和上层应用之间的接口，上层应用比如：串行SCSI 协议，HBA 的驱动提供了FC-4 的接口函数，FC-4 支持多协议，如：FCP-SCSI，FC-IP，FC-VI。



FC与SCSI协议的关系：

- FC通道并不是SCSI的替代，FC可以通过构建帧来传输SCSI的指令、数据和状态信息单元。
- SCSI是位于光纤通道协议栈FC4的上层协议，SCSI是FC协议的子集。

FC存储设备

- 存储设备上的FC接口模块提供了应用服务器与存储系统的业务接口，用于接收应用服务器发出的数据交换命令。

模块电源指示灯

8Gb FC

模块拉手

FC主机端口

FC主机端口 link/speed指示灯

P0 P1 P2 P3

FC主机端口

8Gb FC接口模块提供传输速率为8Gbit/s的主机端口。当连接的设备传输速率低于主机端口速率时，主机端口将自动适应传输速率，以保证数据传输通道的连通性和数据传输速率的一致性。

光纤交换机

- 光纤网络路由直接连接的方式
- 发起者和目标设备建立直接的连接
- 独享光纤的所有带宽
- 区域Zone的划分



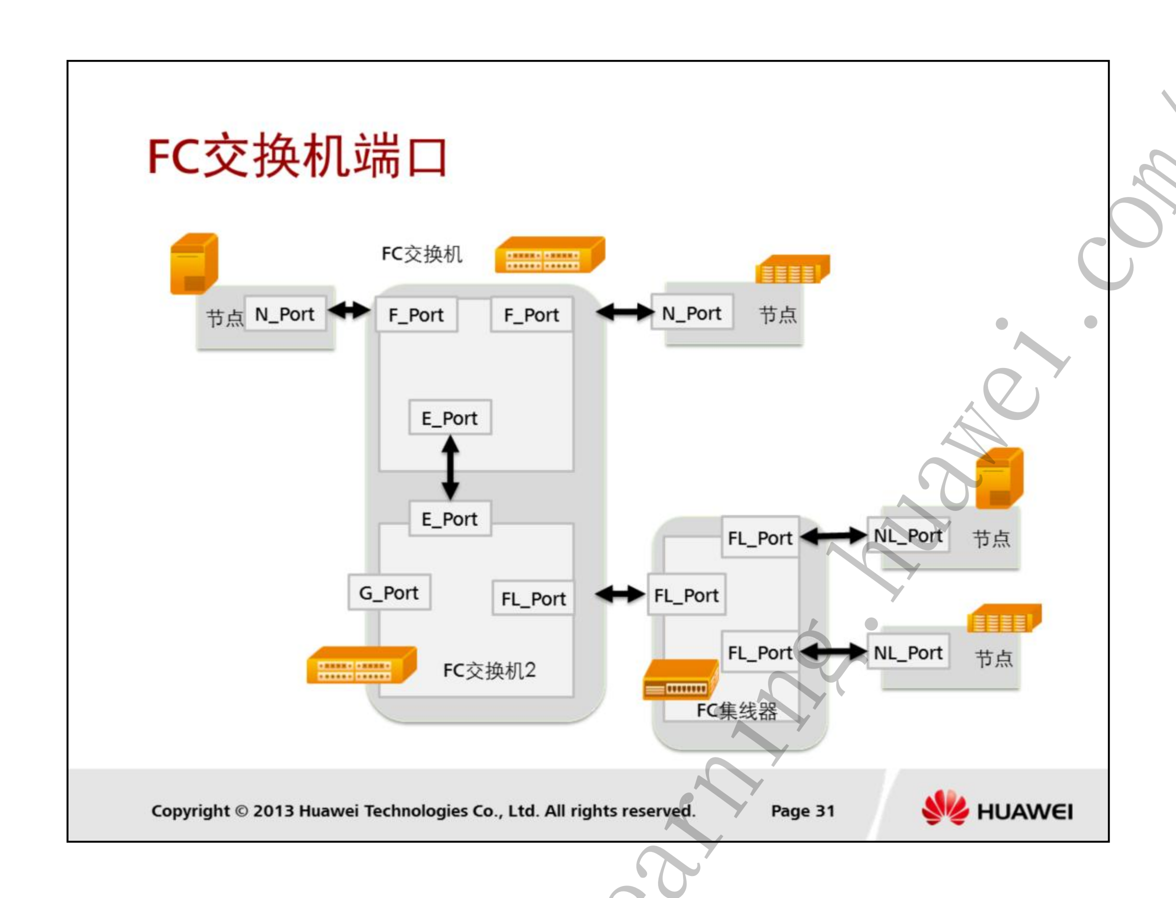
 Page 30 

光纤通道交换机使用光纤网络路由直接连接的方式，发起者和目标设备可以通过光纤通道交换机中的路由软件建立直接的连接以独享光纤的所有带宽。

光纤通道交换机是SAN的核心，它连接着主机和存储设备。一般可分为入门级交换机、工作组级光纤交换机、核心级光纤交换机。工作组级光纤交换机应用最多的领域是小型SAN，通过级联交换机，能够建立一个大型的、虚拟的、具有分布式优点的交换机，并且它可以跨越的距离非常大。核心级交换机（又叫导向器）一般位于大型SAN的中心，支持光纤以外的协议（像InfiniBand）、高级光纤服务（例如：安全性、中继线和帧过滤等），核心光纤交换机往往采用基于"刀片式"的热插拔电路板。

与以太网交换机相比，FC交换机用于构建光纤网络，而以太网交换机用于构建以太网络；光纤交换机中使用的是FC协议，以太网交换机使用的是TCP/IP协议。

光纤交换机上为了不同设备之间的访问隔离引入了zone的概念，zone的功能类似于以太网交换机上的VLAN功能，它是将连接在SAN网络中的设备（主机和存储），逻辑上划到为不同的区域内，使得不同区域中的设备相互间不能通过FC网络直接访问，从而实现网络中的设备之间的相互隔离。



光纤网交换机中具有一些称为端口（Port）的连接部件，不同的端口根据其所连接的设备类型，所起到的作用是不同的。光纤通道标准定义了以下端口：

- F_Port

F端口也被称为光纤网端口，用于将服务器和存储设备连接到交换机上。一个被连接到交换机F端口的设备就是一个节点（Node），用光纤通道术语来说，它被看作是一个N端口（N_Port）。如果是在仲裁环路拓扑结构中，则被称为NL端口（NL_Port）。交换机通过特定的N端口或者是NL端口名称来识别这些光纤通道节点。

- E_Port

E端口也被称为扩展端口，被用于进行交换机之间的连接。

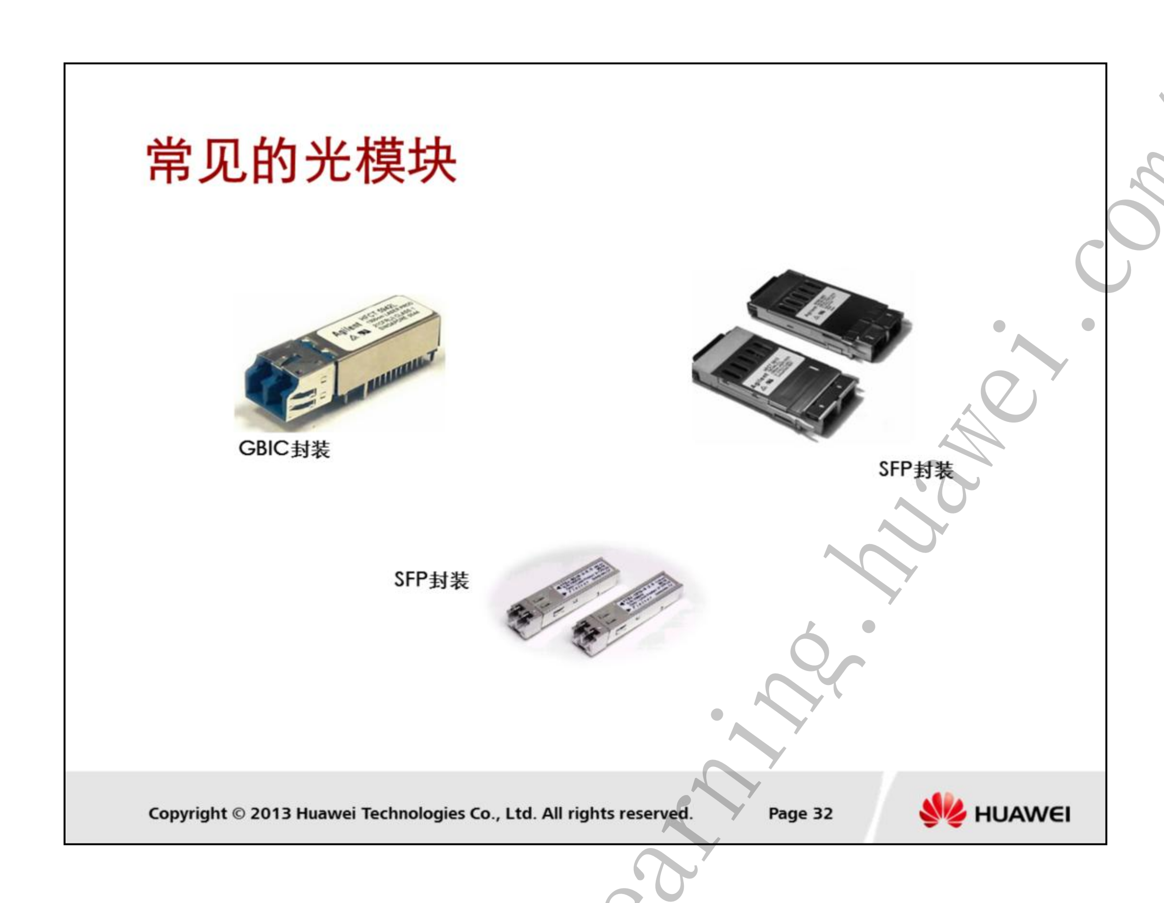
- FL_Port

FC交换机的一个交换端口可以作为环路的组成部分，数据可以从交换机中传输到环上。在环路环境下正常工作的一个交换端口称之为 “FL_port”。

- G_Port：通用端口G端口

G端口是一种通用的端口，根据具体的实现方案，可以作为F端口或是E端口使用，也就意味着G端口实际上可以被用作两种端口功能的组合。由于G端口的自适应性，在进行FC-SAN的多交换机配置环境时，G端口为交换机提供了更好的灵活性并降低了每个端口所耗费的管理成本。

目前光纤交换机支持的端口速率有1 、2 、4、8 Gb/s 。



光通道交换机光模块由光电子器件、功能电路和光接口等组成。光电子器件包括发射和接收两部分。

- 按照速率分：以太网应用的100Base（百兆）、1000Base（千兆）、10GE SDH应用的155M、622M、2.5G、10G
- 按照封装分：1×9、SFF、SFP、GBIC、XENPAK、XFP，
- 按照光纤的类型分：单模光纤连接器、多模光纤连接器；
- 按照光纤连接器的连接头形式分：FC，SC，ST，LC，MU，MTRJ 等等，目前常用的有FC，SC，ST，LC。

在华为存储设备中，采用多模的LC连接器。

常用光纤连接介质

介质类型	发射器	速率	距离
9μm单模光纤	1550nm长波光激光器	1Gb/s	2m – 50Km
		2Gb/s	2m – 50Km
	1300nm长波光激光器	1Gb/s	2m – 10Km
		2Gb/s	2m – 2Km
		4Gb/s	2m – 2Km
50μm多模光纤	850nm短波光激光器	1Gb/s	0.5m-500m
		2Gb/s	0.5m-300m
		4Gb/s	0.5m-170m
62.5μm多模光纤		1Gb/s	0.5m-300m
		2Gb/s	0.5m-150m
		4Gb/s	0.5m-70m



- 按照光纤的模式分类：
  - 单模光纤（Single Mode Fiber）
  - 多模光纤（Multi Mode Fiber）

HBA卡

- HBA（Host Bus Adapter）：
  - 主机总线适配器，就是连接主机I/O总线和计算机内存系统的I/O适配器。
- 分类：
  - FC HBA、SCSI HBA、SAS HBA、iSCSI HBA等。
- 用途：
  - 用于服务器、海量存储子网络、外设间通过集线器、交换机和点对点连接进行双向、串行数据通讯。

各类HBA卡



 Page 34



FC HBA卡是联系服务器与存储区域网络的设备。与网络接口卡（Network Interface Card，NIC）类似，HBA提供了服务器协议和光纤网络交换机之间进行转换的能力。HBA连接到服务器的PCI总线，通过软件驱动程序来提供对光纤通道网络的支持。HBA根据厂商的实现，可以使用单端口或者多端口配置。在多端口环境中，数据流拥有额外的数据路径，用于通过单个HBA在服务器和光纤网交换机之间传输数据。一个HBA可以拥有多个端口，而一个服务器也可以使用多个HBA，这样的配置更具灵活性，可以实现到多个连节点的单独连接，也能实现到同一连节点的路径冗余以避免单点故障的风险。

FC HBA的主要厂家有Emulex、Qlogic、LSI、JNI（已经并入AMCC，LSI存储兼容列表写的HBA为AMCC）、Agilent、Adaptec、IBM、HP、SUN。

FC HBA卡WWN

- WWNN（World Wide Node Name）
- WWPN（World Wide Port Name）

WWPN
WWNN
输入
输出
链接
FC网络
• P2P
• FC-AL
• FC-SW

 Page 35 HUAWEI

FC HBA的WWN具有两种类型：

- WWNN（World Wide Node Name）

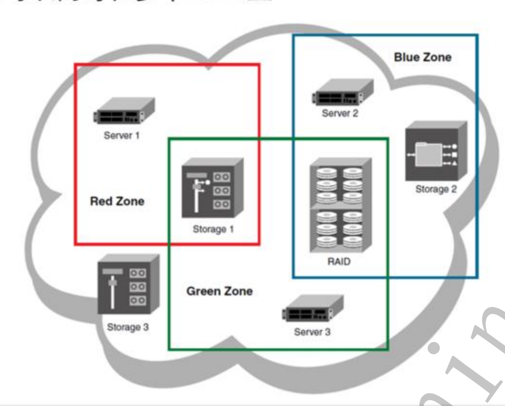
全球唯一节点名字，由光纤通道进行映射的分配给每一个上层节点一个全球惟一的64位标识符，一个HBA上的所有端口共享一个WWNN。在计算机处理中，一个WWNN被分配给一个接入到光纤网中的节点（一个端点，如，一个设备）。WWNN可以被一个或者多个不同的端口（每个端口拥有不同的WWPN，并且属于同一个节点）共同使用。

- WWPN（World Wide Port Name）

全球惟一端口名字，分配给每一个光纤通道端口的全球惟一的64 位标示符，每个WWPN被该端口独享。WWPN在存储区域网络中的应用就等同于MAC地址在以太网协议中的应用。

FC交换机Zone概念

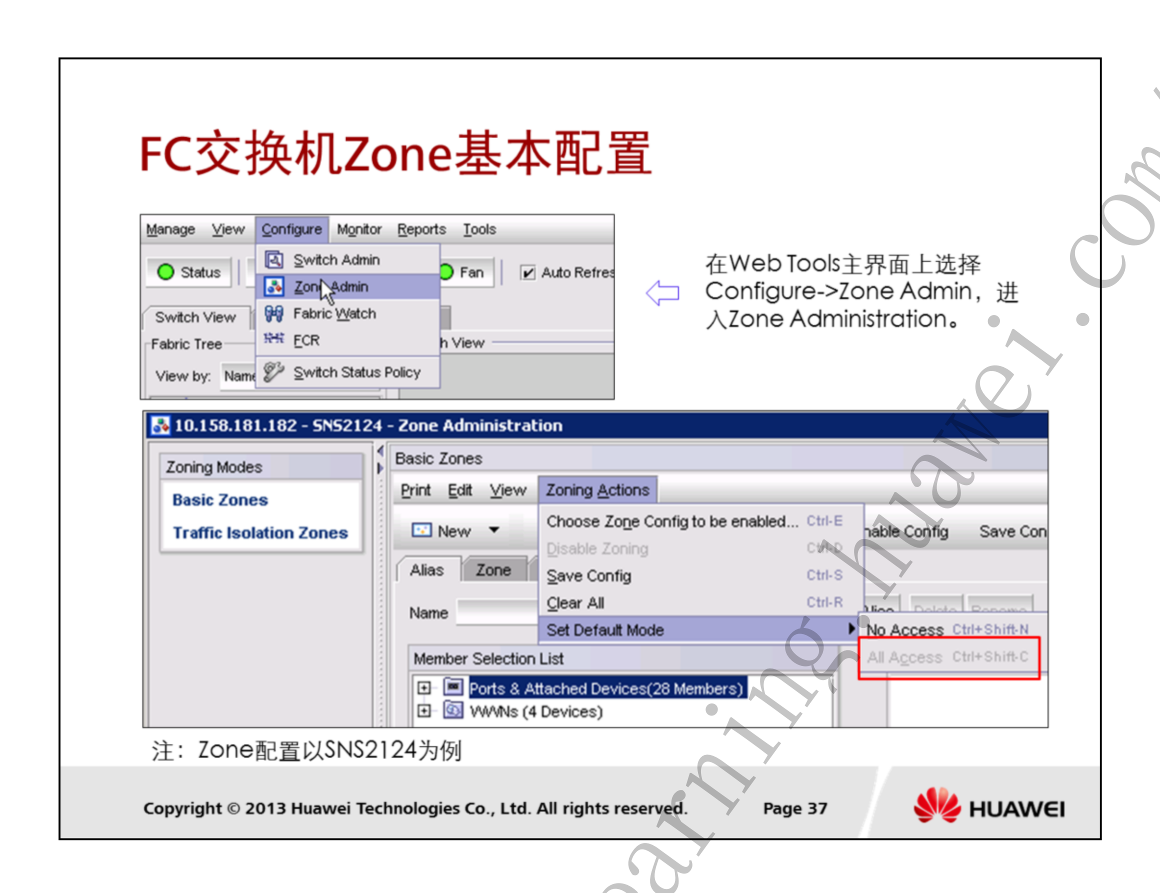
- Zone是可进行互通的端口或设备的名称构成的集合
- 在一个zone里的设备只能与同一个zone中的其他设备通信
- 一个设备可以同时在多个zone里



 Page 36



- Basic Zones:（基本分区）
  - 设置基本分区，控制各设备或端口之间的访问权限。
- Traffic Isolation Zones: （流量隔离分区）
  - 当存在多条ISL级联链路（多个E_Port）时，用于指定某条ISL链路只负责传送与该链路E_Port在同一Zone内的端口的流量。



- 设置默认Zone模式：

  1. 打开“Zone Administration”页面。
  2. 选择“Zoning Actions > Set Default Mode”，选中All Access,允许Fabric内所有设备可以相互访问。

创建Alias

Basic Zones
Print Edit View Zoning Actions
Step 3
New Resource View Refresh Enable Config Save Config Clear All
Alias Zone Zone Config
Step 1
Name S3900 New Alias Delete Rename
Member Selection List
Ports & Attached Devices(50 Members)
1(SNS2248)(48 Ports)
WWNs (2 Devices)
Emulex Corporation 20:00:00:00:c9:87:f7:c3
[28] "HUAWEI S3900-M200 2105" 21:00:00:22:a1:09:c8:28
Step 2
Alias Members
1 Member.
1,0(port0,F-Port,PID: 010000)

 Page 38 HUAWEI

- 创建Alias的主要步骤：
  1. 点击“New Alias”，在弹出的对话框中，输入名字，点击“OK”；
  2. 在“Member Selection List”中选择需要加入Alias的端口，点击 ▷ ，将端口加入Alias Members；
  3. 选择“Save Config”，保存配置。

创建Zone

Step 3
Step 1
Step 2

 Page 39 HUAWEI

- 创建Zone的主要步骤：
  1. 点击“New Zone”，在弹出的对话框中，输入名字，点击“OK”；
  2. 在“Member Selection List”中选择需要加入Zone的端口，点击 ▷ ，将端口加入Zone Members；
  3. 选择“Save Config”，保存配置。

创建并激活Zone Config

Page 40

HUAWEI

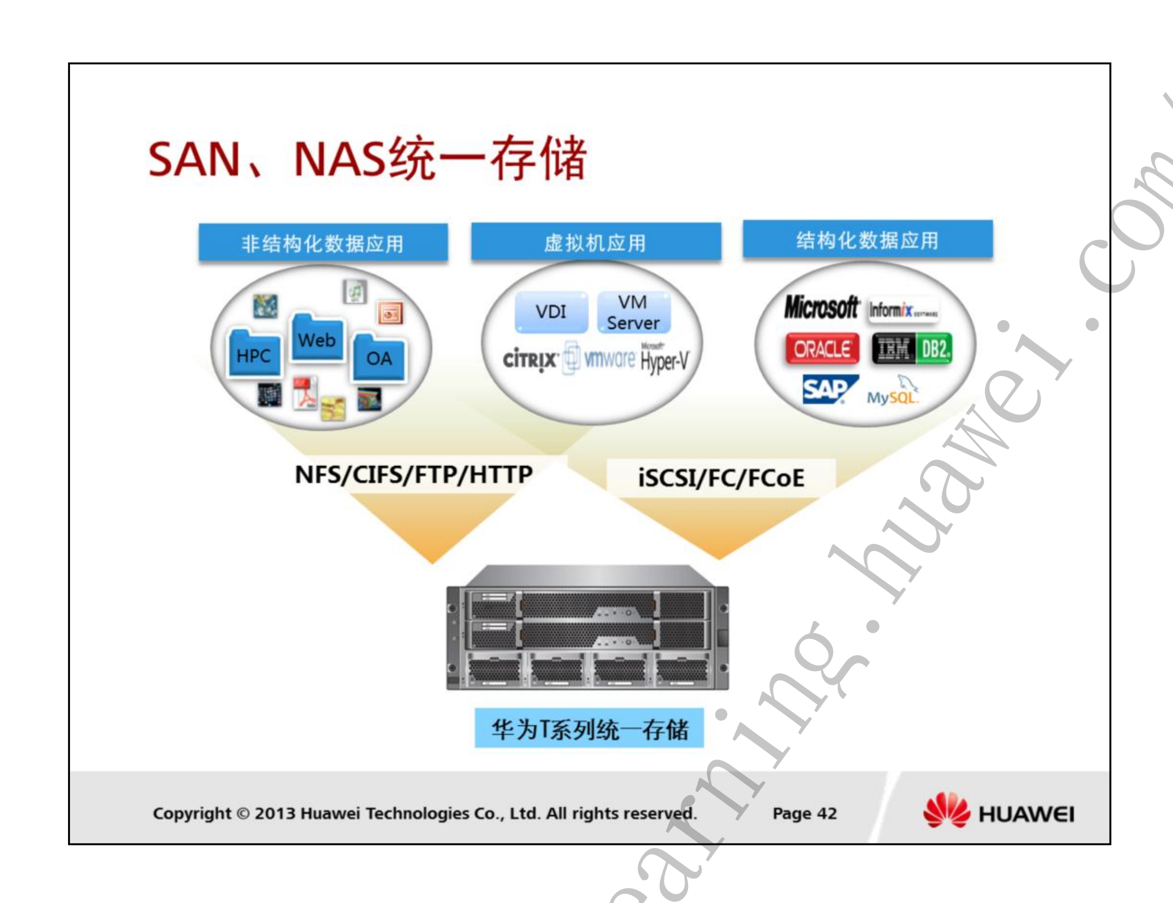
- 创建Zone Config的主要步骤：
    1. 点击“New Zone Config”，在弹出的对话框中，输入名字，点击“OK”；
    2. 在“Member Selection List”中选择需要配置的Zone，点击 ；
    3. 选择“Save Config”，保存配置。
    4. 选择“Enable Config”，激活配置。

创建Zone Config后需要为Zone Config命名。



目录
1. DAS存储基础
2. SCSI协议介绍
3. SAN存储基础
4. FC连接与协议
5. 华为SAN存储应用
Copyright © 2013 Huawei Technologies Co., Ltd. All rights reserved.
Page 41
HUAWEI

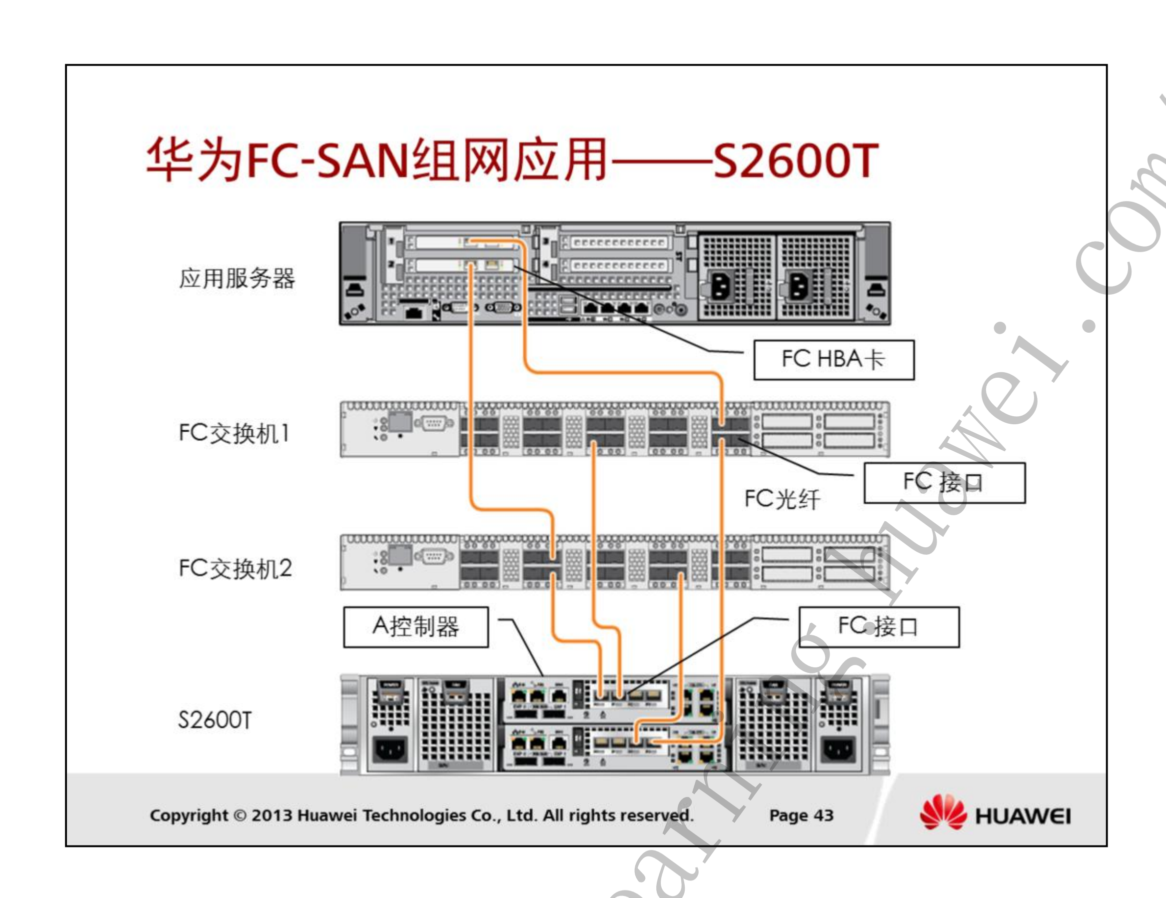




华为T系列存储设备实现了SAN和NAS的统一，满足用户对SAN存储和NAS存储应用需求。其主要应用场景为用户有FC SAN、 IP SAN,及NAS统一存储需求 。T系列存储设备主要的优势包括:

- 用户不用去单独规划SAN和NAS的存储容量
- 用户可以自由分配存储来满足应用环境的需要
- 用户不用关心对数据块及文件存储容量利用率的问题

注：IP SAN和NAS相关内容将在后面章节详细介绍。

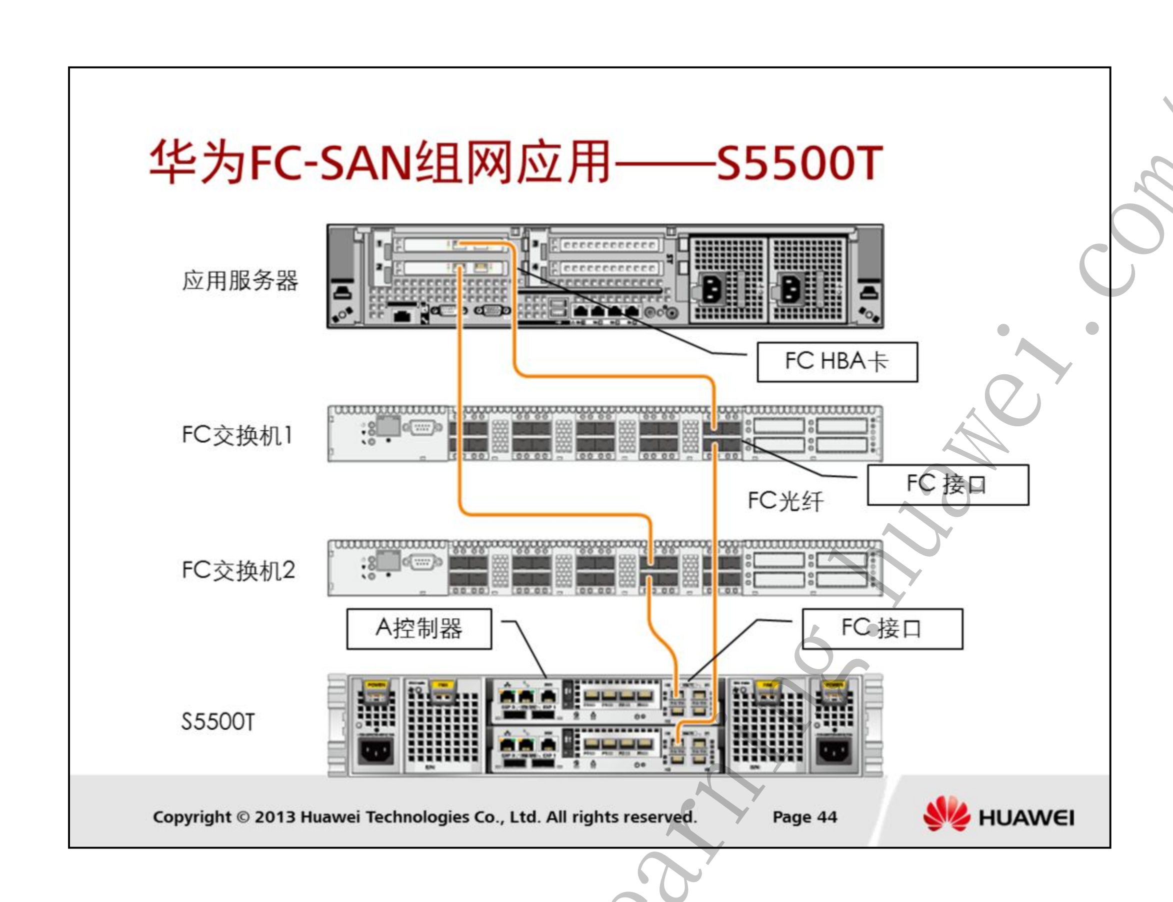


华为S2600T系列存储阵列设备可以通过光纤将控制框FC主机端口与应用服务器端口进行连接，建立业务通道，实现存储设备与应用服务器之间的通信。

存储设备与应用服务器的连接方式有两种： 直接连接和通过FC光纤交换机连接。

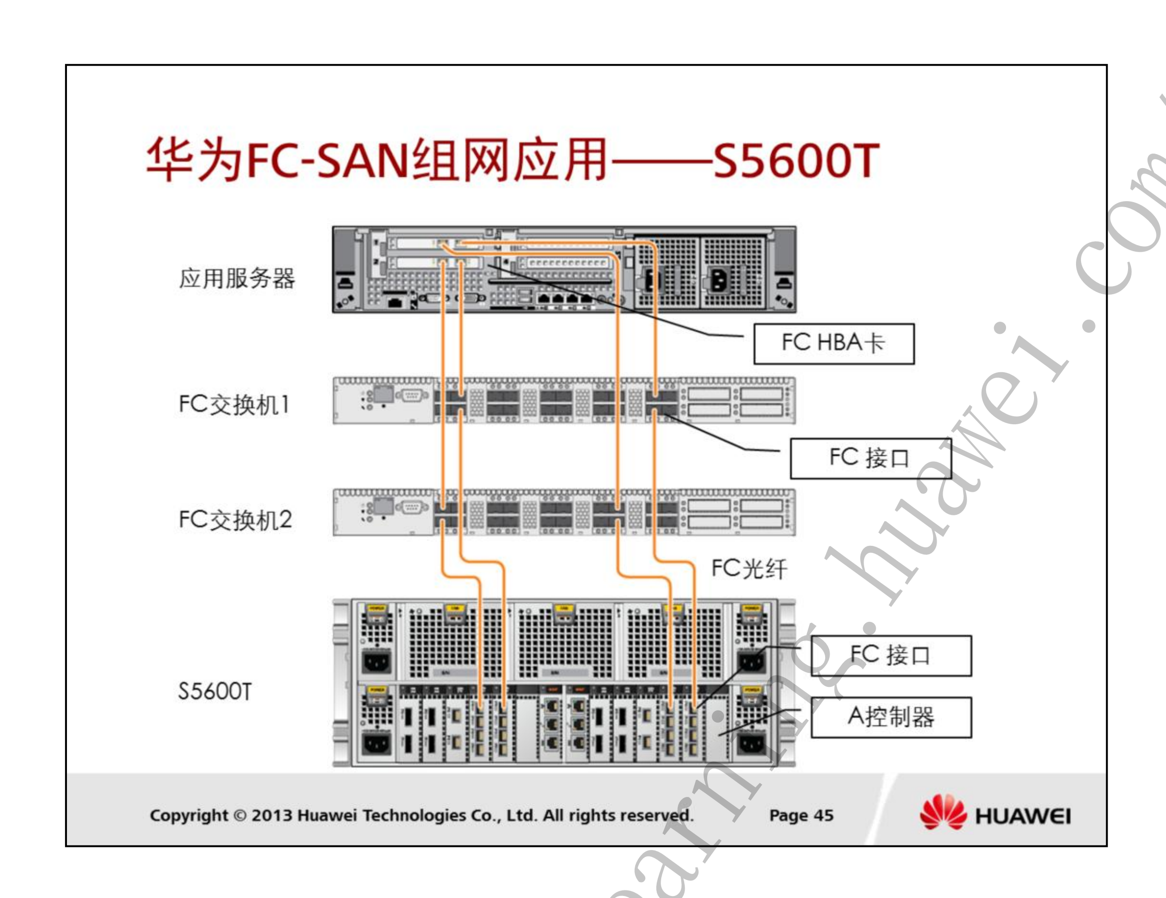
通过FC光纤交换机连接时，需要提前做好以下规划：

- 规划交换机Zone划分。确定划分的Zone对应的交换机端口。
- 交换机的首尾端口常用于交换机之间的组网，不建议用于主机服务器与存储阵列的连接。



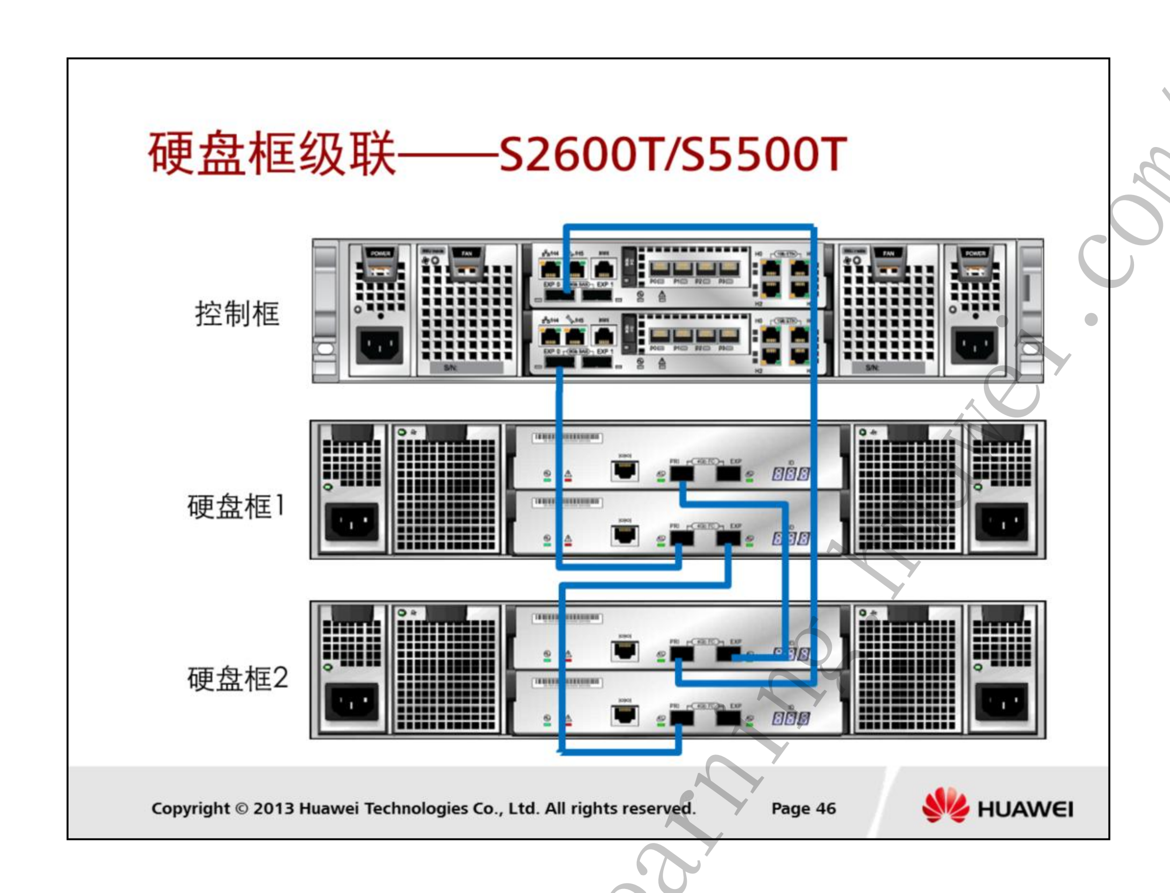
华为S5500T存储阵列设备控制框可以通过8Gb FC主机端口连接到应用服务器。

与S2600T存储设备一样，S5500T存储阵列与应用服务器的连接方式有两种： 直接连接和通过FC光纤交换机连接。同样，通过FC光纤交换机连接时，需要提前做好交换机zone规划。



华为S5600T存储阵列设备控制框为4U盘控分离设计，可以通过8Gb FC主机端口连接到应用服务器。控制框上的接口模块都是冗余配置，因此，建议将控制框冗余连接至应用服务器。

与S2600T存储设备一样， S5600T存储阵列设备与应用服务器的连接方式有两种： 直接连接和通过FC光纤交换机连接。同样，通过FC光纤交换机连接时，需要提前做好交换机zone规划。



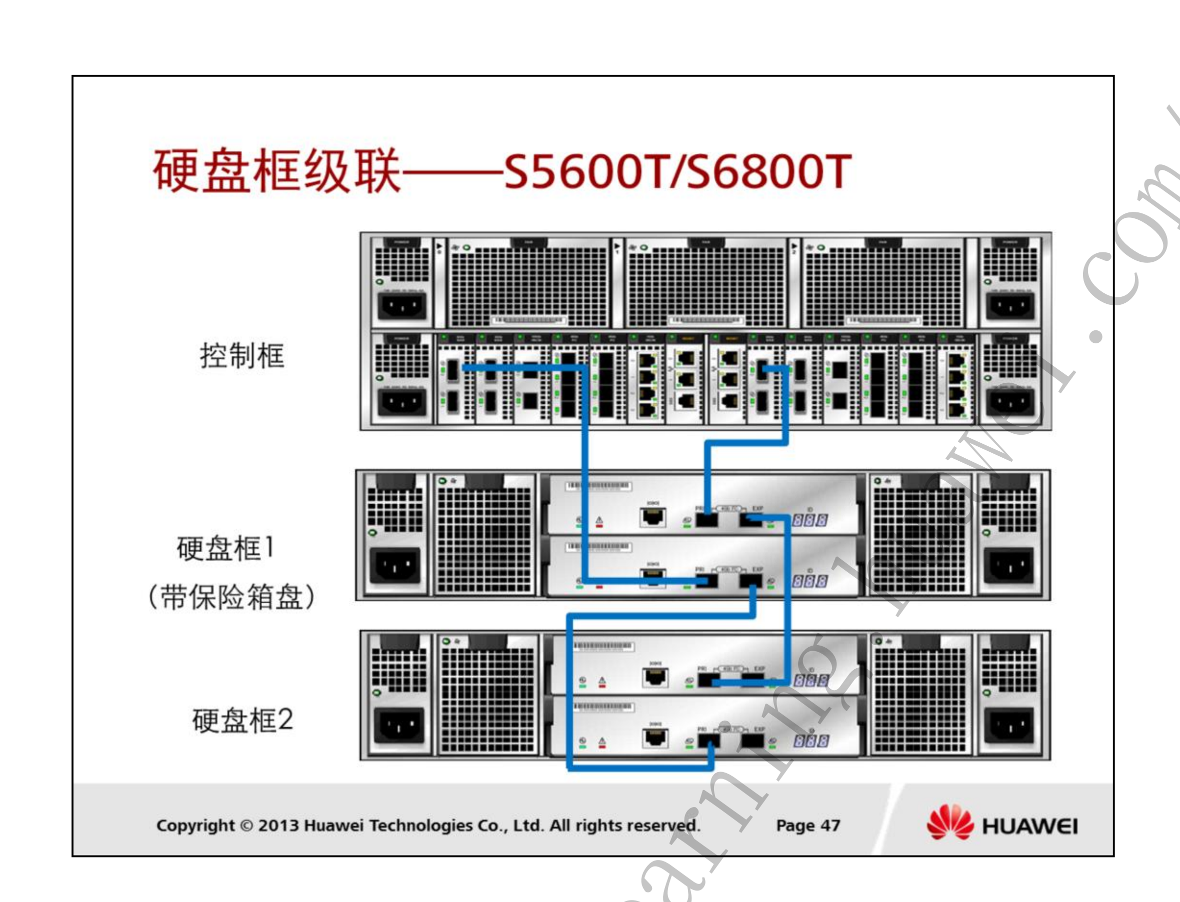
华为T系列存储设备可通过mini SAS线缆实现控制框与硬盘框以及硬盘框之间的级联。

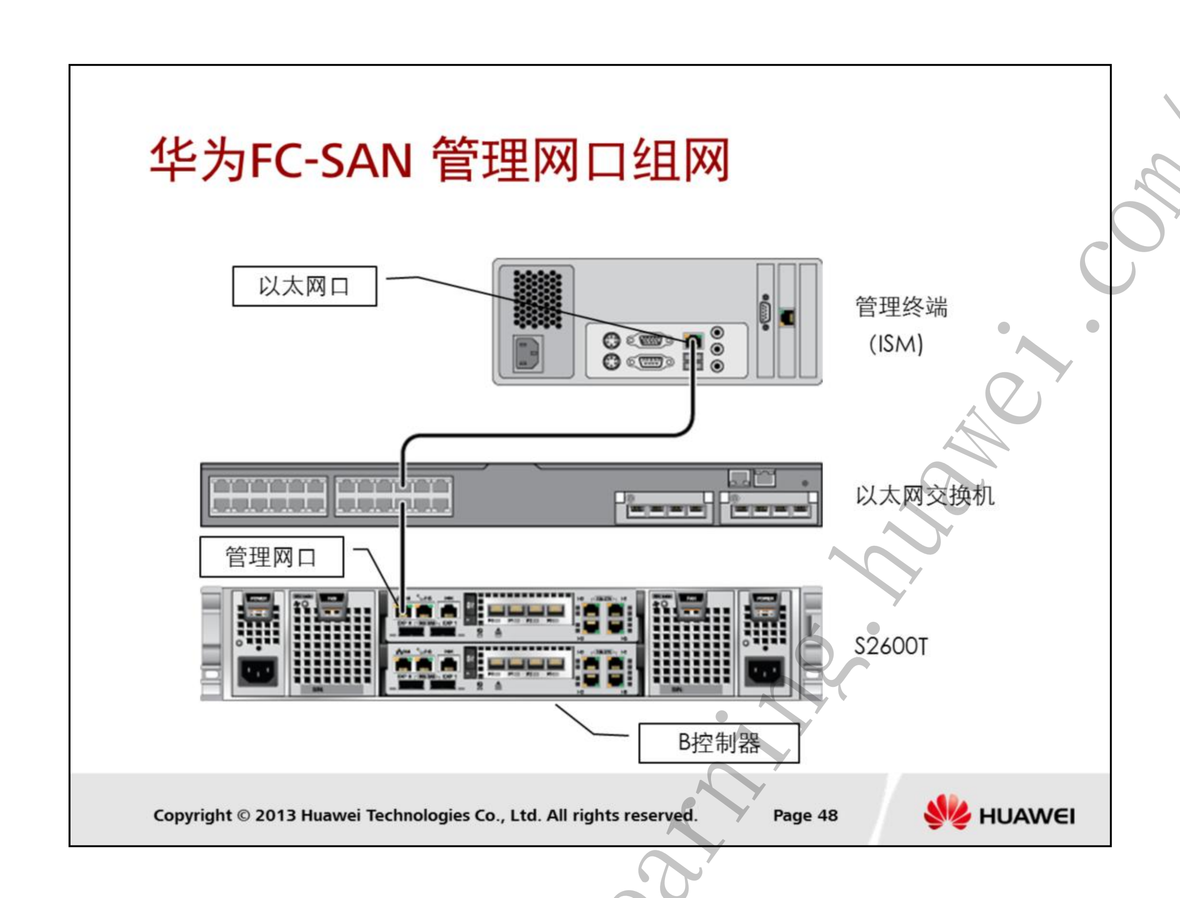
级联硬盘框时，应采用正反向冗余连接的方式。正向连接和反向连接的描述如下：

- 正向连接：假设控制框级联了1、2、3三个硬盘框，控制器A上的级联端口连接1号硬盘框级联模块A上的PRI级联端口，1号硬盘框级联模块A上的EXP级联端口又连接2号硬盘框级联模块A上的PRI级联端口。以此类推，将控制器A和所有硬盘框级联模块A上的级联端口依次连接的方法称为正向连接。
- 反向连接：假设控制框级联了1、2、3三个硬盘框，且控制器A和所有硬盘框级联模块A上的级联端口依次正向连接。控制器B上的级联端口连接3号硬盘框级联模块B上的PRI级联端口，3号硬盘框级联模块B上的EXP级联端口又连接2号硬盘框级联模块B上的PRI级联端口。以此类推，将控制框B和所有硬盘框级联模块B上的级联端口依次连接的方法称为反向连接。

级联硬盘框之前，需要遵循下面的级联规则进行连接：

- 存储设备上的所有EXP级联端口只能与PRI级联端口相连。
- 如果级联两个或两个以上数量的硬盘框，建议根据控制框上级联端口的数量组建多个级联环路，每个级联环路上的硬盘框数量尽量保持一致。
- 保险箱所在的硬盘框需正向连接。
- 保险箱盘所在的框必须接在0环路上（即控制框0号槽位的SAS卡的P0口上）。


硬盘框级联——S5600T/S6800T
控制框
硬盘框1
（带保险箱盘）
硬盘框2
Copyright © 2013 Huawei Technologies Co., Ltd. All rights reserved.
Page 47
HUAWEI

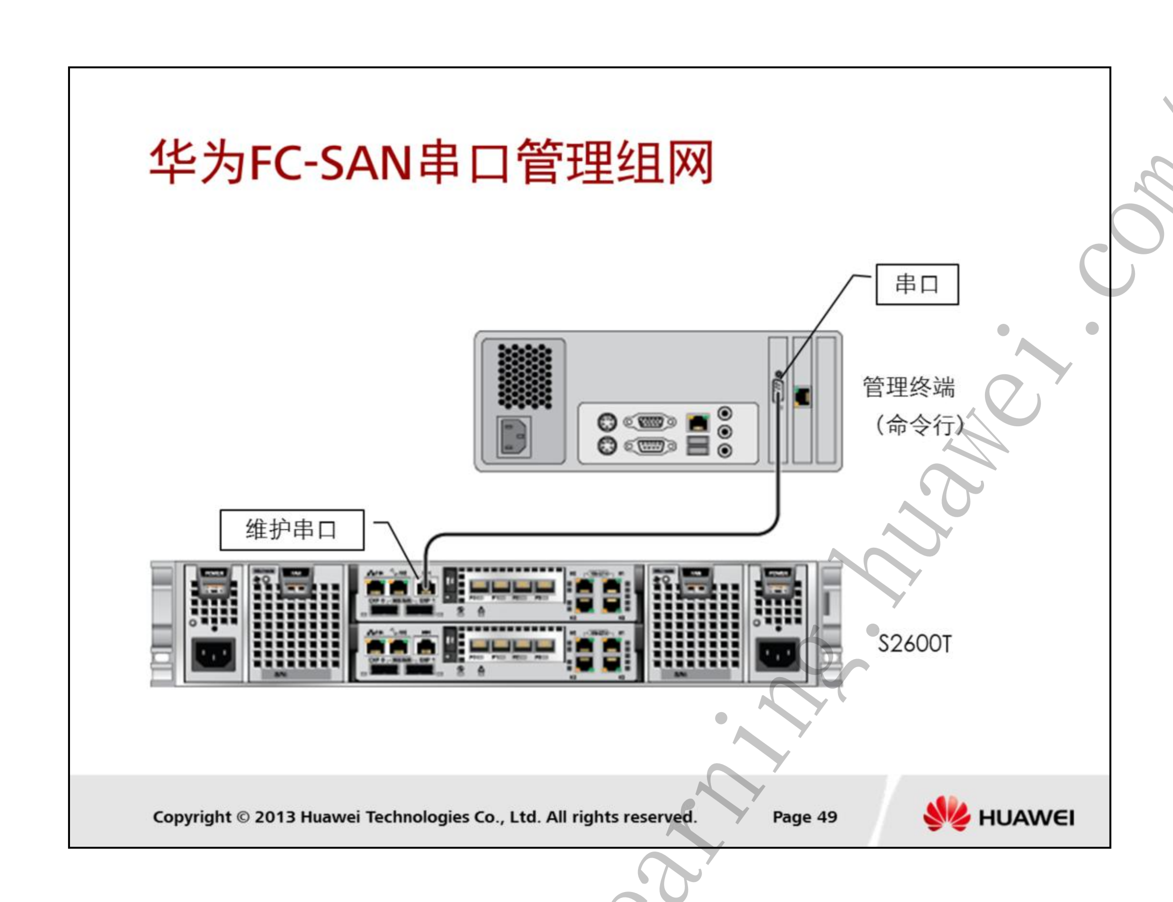


华为存储设备可以通过网线将控制框管理网口与维护终端网口进行连接，建立管理通道，实现维护终端对存储设备的管理和维护。每个控制器上都设置有一个管理网口和维护网口。维护网口用于紧急情况下的特殊维护，因此一般只使用管理网口来进行配置管理。

存储设备与维护终端的连接方式有两种：

- 直接连接。
- 通过以太网交换机连接

通过以太网交换机连接方式时，需要保持管理终端与存储阵列设备管理网口间的通讯正常。



华为T系列存储设备的串口和维护终端之间通过串口线缆连接后，可以通过串口对存储系统进行管理和维护。有两种端口类型的串口线缆：RJ-45转DB9串口线缆和双RJ-45串口线缆。两种类型串口线缆的选用视维护终端的串口类型而定。一般而言，存储系统通过RJ-45转DB9串口线缆连接。



总结

- DAS存储基础知识
- SCSI协议
- SAN存储的基础知识
- FC连接和协议封装
- 华为SAN存储的应用





思考题

- DAS存储的局限性有哪些?
- 如何描述SCSI协议模型？SCSI协议如何实现寻址操作?
- SAN存储典型组网有哪几种？各自的组网特点是什么?
- SAN存储网络由哪几个主要组件组成?
- FC主要的拓扑结构包括哪几种？各自有什么特点?
- FC协议分为几层，各层的主要功能和特性是什么？与SCSI协议的关系是什么?
- 华为SAN存储设备有哪些应用组网，如何实现组网连接?





习题

- 判断题
  1. SCSI 采取了分层结构，主要包括三层，即SCSI应用层，SCSI传输层和SCSI互连层。(T or F)
- 多选题
  1. FC光纤交换机包括的主要端口类型是:
     A. F_Port
     B. N_Port
     C. E_Port
     D. G_Port

 Page 52



习题答案:

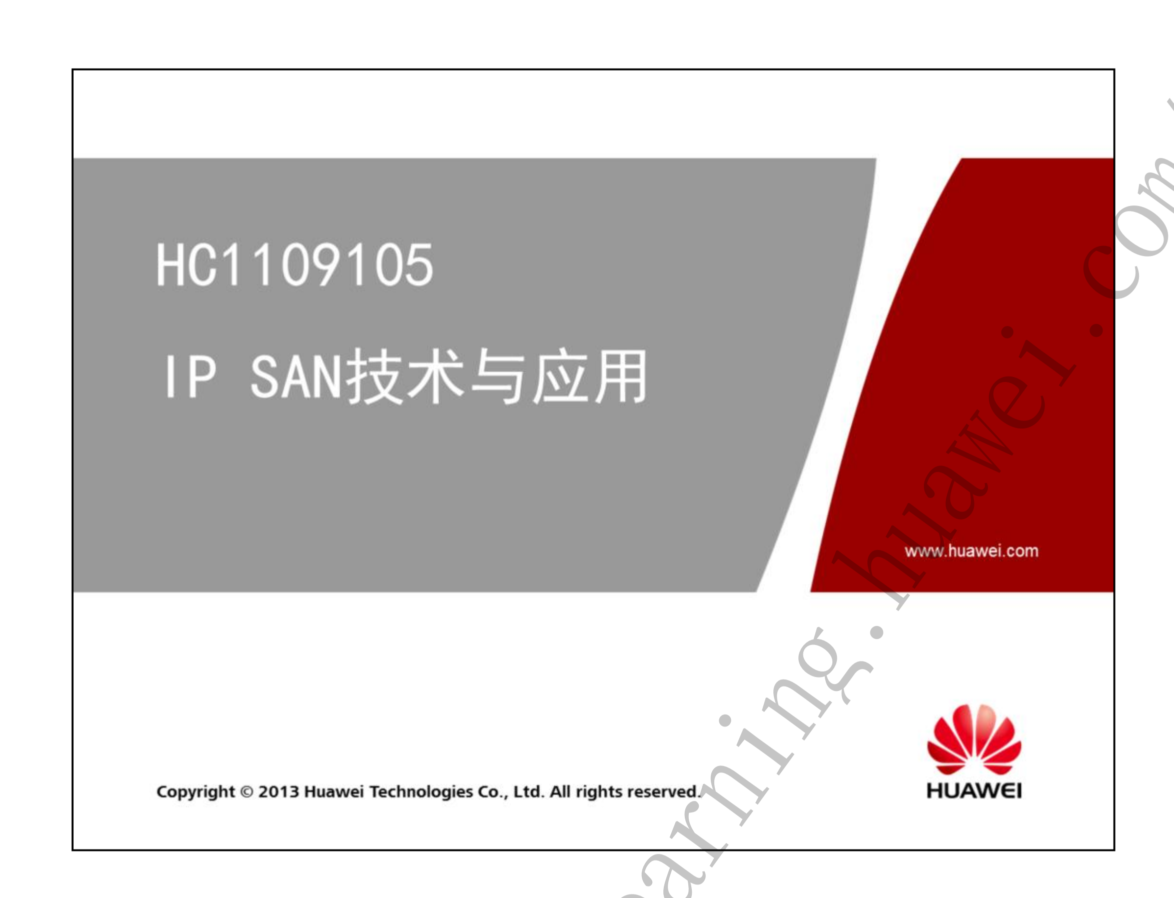
- 判断题：1.T
- 多选题：1.ACD

Thank you

www.huawei.com

HC1109105 IP SAN 技术及应用




HC1109105
IP SAN技术与应用
www.huawei.com
Copyright © 2013 Huawei Technologies Co., Ltd. All rights reserved.
HUAWEI



目标

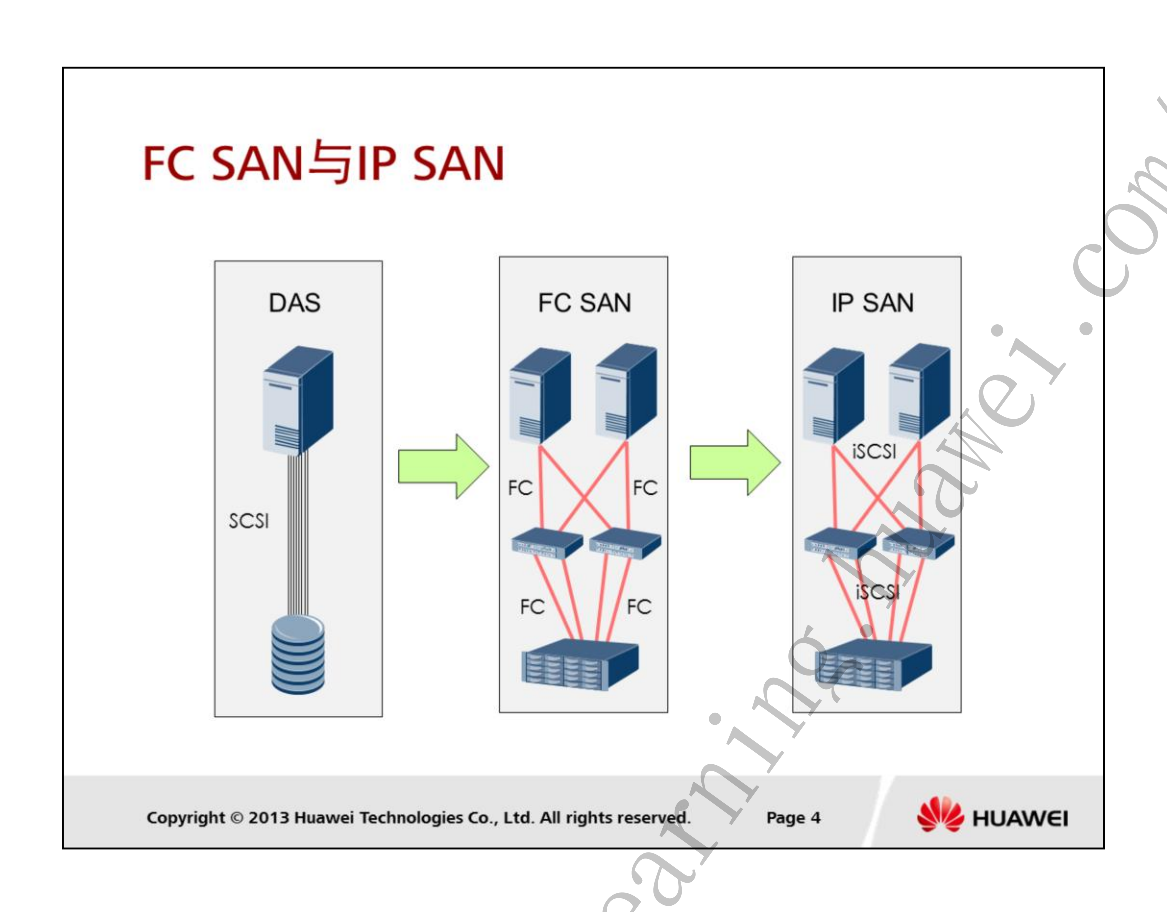
- 学习完本章节后，您将能够：
  - 了解IP SAN产生与发展的背景
  - 掌握IP SAN组成和组网连接
  - 理解iSCSI协议
  - 理解FC协议与TCP协议融合
  - 掌握华为IP SAN存储的实现与应用





目录

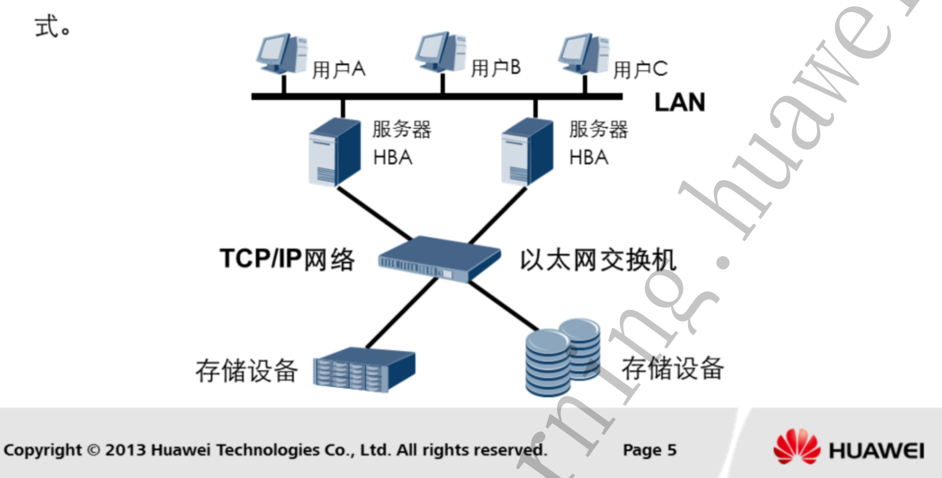




前面介绍了基于FC协议的FC SAN，主要应用于对于性能、冗余度和可获得性都有较高要求的中高端存储需求。由于其高昂的价格、技术和配置复杂、有限的传输距离、不同厂商设备互联共享等问题，也限制了其使用的范围。为了提高SAN的普及程度和应用范围，并充分利用SAN本身所具备的架构优势，SAN的发展方向开始考虑和已经普及的并且相对廉价的IP网络进行融合。

基于TCP/IP协议的以太网的IP-SAN存储开始进入人们的视野，并得到了网络厂商的广泛支持。与基于光纤通道技术的FC-SAN相比较而言，IP-SAN存储系统具有节约大量成本、加快实施速度、优化可靠性以及增强扩展能力等优点。

简单而言，IP SAN存储就是使用IP网络而不使用光纤网络来做服务器和存储设备的连接技术。IP SAN存储是基于IP网络来实现块级数据存储的方式。目前除了标准已获通过的iSCSI，还有FCIP、iFCP等标准。其中iSCSI发展最为迅速，已经成为IP存储的中流砥柱。基于iSCSI的SAN的目的就是要使用本地iSCSI Initiator（启动器，通常为服务器）通过IP网络和iSCSI Target（目标器，通常为存储设备）来建立SAN网络。



IP SAN是标准的TCP/IP协议和SCSI指令集相结合的产物，以其协议标准化、整体成本低廉和维护简便等优势成为网络存储领域的重要产品形态。

IP SAN是基于IP网络来实现数据块传输的网络存储形态，与传统FC SAN的最大区别在于传输协议和传输介质的不同。目前常见的IP SAN协议有iSCSI、FCIP、iFCP等，其中iSCSI是发展最快的协议标准，大多时候人们所说的IP SAN就是指基于iSCSI实现的SAN。

IP SAN把SCSI指令集封装在了TCP/IP上。这就好比，不管我们是选择哪家快递公司，最终都是把我们想要发送的东西发送至目的地，都是由我们发起寄送请求，快递公司进行响应，差别只在于快递公司不同而已。iSCSI则是全新建立在TCP/IP和SCSI指令集的基础上的标准协议，所以其开放性和扩展性更好。这也是其大行其道的原因。



IP SAN主要基于iSCSI实现。iSCSI协议是建立在TCP/IP协议和SCSI指令集的基础之上的标准化协议。正是其优良的基因决定了其巨大的优势。

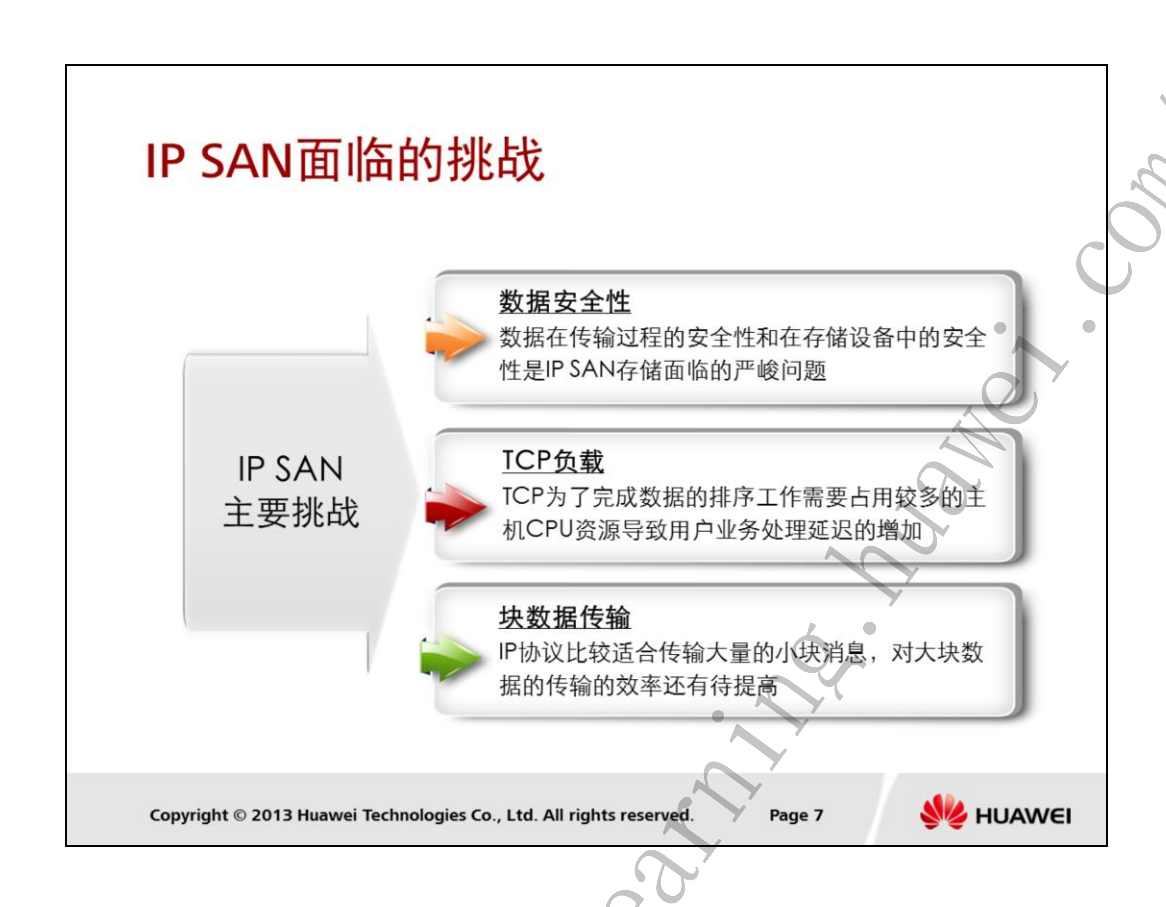
那么，TCP/IP和SCSI有什么优势呢？

第一，它们都是标准化协议，所以有大量的标准化设备可供采用；

第二，它们都是已经发展多年的成熟性协议，具有广泛的群众基础；

第三，作为标准在各类软件硬件开发中已经广泛采用。

IP SAN全盘继承了父母的优良基因，从而具备了很多方面的优势。通过这些优势，给客户带来了哪些好处呢？最重要的一点就是IP SAN总体拥有成本（TCO）低，非常有利于其广泛的应用和推广。总体拥有成本（TCO）是包含很多内容的，比如建设一个存储系统，则需要购买磁盘阵列、接入设备（HBA和交换机）、需要人员培训、日常维护、后续扩容、容灾扩展等。IP SAN因为IP网络的广泛应用优势，可以大幅降低单次采购的接入设备采购成本、减少维护成本，而且后续扩容和网络扩展成本也大幅降低。



尽管IP存储标准早已建立且应用，但是将其真正广泛应用到存储环境中还需要解决几个关键问题：

数据安全性：企业网络中最重要的还是数据，所以，SAN中保存的数据的安全性和可靠性应当受到格外的重视。传统的FC-SAN由于FC网络的异构性，和传统的业务IP网络从物理上隔离，从而保证了在SAN中传输和存储的数据安全性。然而，当存储设备通过IP架构进行远程连接时，尽管IP协议可以应用IPSec以保障数据的安全性，但也只能提供数据在网络传输过程的动态安全性，并不能保证数据被保存在存储设备上的静态安全性。由于IP网络是开放式网络，仍然存在众多安全漏洞，并且，使用IP网络构建的IP-SAN和传统的IP业务很难从物理上完全隔离。所以，提高数据在传输过程的安全性和在存储设备中的安全性是IP存储面临的严峻问题。

TCP负载空闲引擎：由于IP协议是无连接不可靠的传输协议，数据的可靠性和完整性是由TCP协议来提供的。而TCP为了完成数据的排序工作需要占用较多的主机CPU资源导致用户业务处理延迟的增加。所以，iSCSI可以采用一种被称为TCP卸载引擎（TCP Off-loading Engine，TOE）的设备，将原本需要由CPU处理完成的TCP封装和解封装过程下移至TOE网卡完成，使CPU更专注于用户事务请求处理和数据包本身的处理，可以有效的降低主机CPU的负载，从而提升CPU的处理性能。

块数据传输问题：FC存储协议具有速率高、延迟低的特点，适合传输大块的数据（Block Data）；而从网络协议上来看，IP协议传输速率相对较低、延迟较高，比较适合传输大量的小块消息。并且，FC在传输数据时将数据封装为2K左右的数据帧进行传输，而以太网则将数据封装为1.5K的数据包在IP网络中进行传递，所以IP协议对大块数据的传输的效率还有待提高。

FC SAN与IP SAN比较

描述	FC SAN	IP SAN
网络速度	1Gb、2Gb、4Gb、8Gb	1Gb、10Gb
网络架构	单独建设光纤网络和HBA卡	使用现有IP网络
传输距离	受到光纤传输距离的限制	理论上没有距离限制
管理、维护	技术和管理较复杂	与IP设备一样操作简单
兼容性	兼容性差	与所有IP网络设备都兼容
性能	非常高的传输和读写性能	目前主流1Gb，10Gb正在发展
成本	购买（光纤交换机、HBA卡、光纤磁盘阵列等）、维护（培训人员、系统设置与监测等）成本高	与FC－SAN相比，购买与维护成本都较低，有更高的投资收益比例
容灾	容灾的硬件、软件成本高	本身可以实现本地和异地容灾，且成本低
安全性	较高	较低

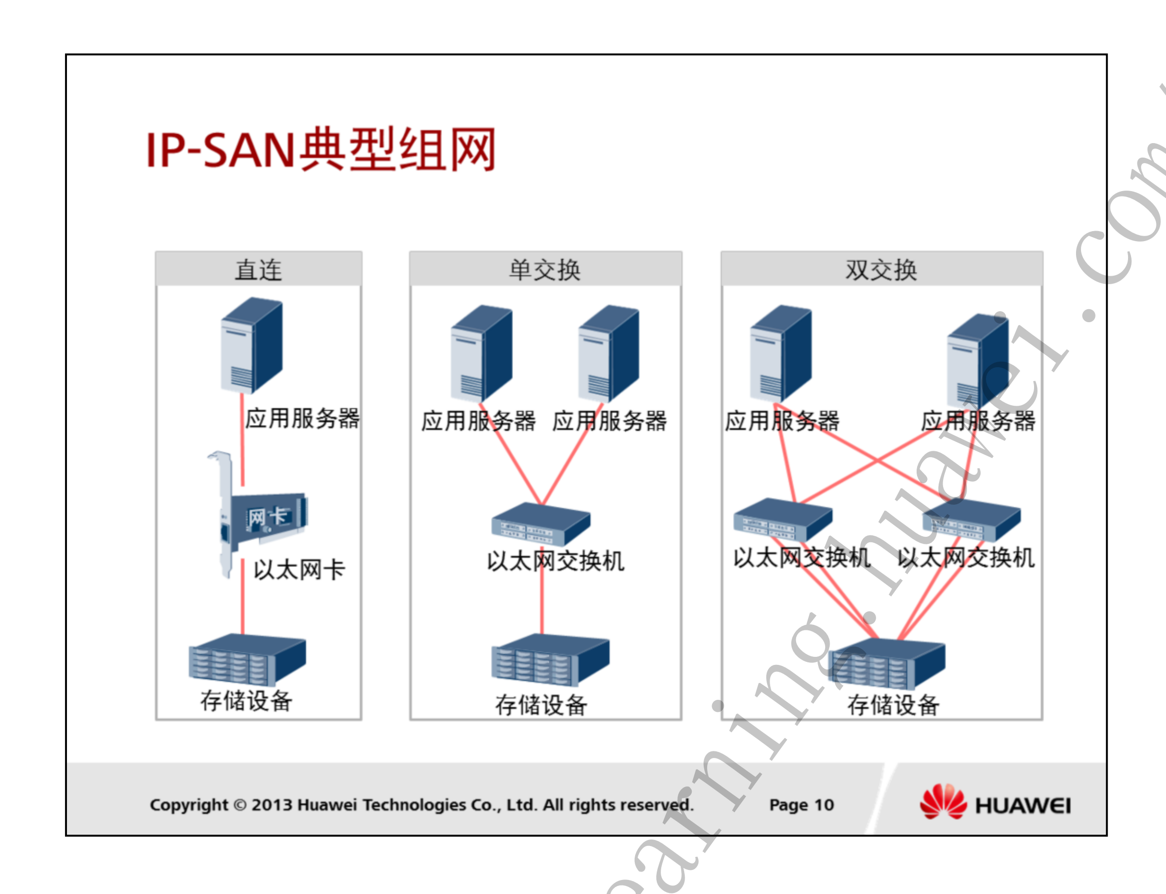
Page 8



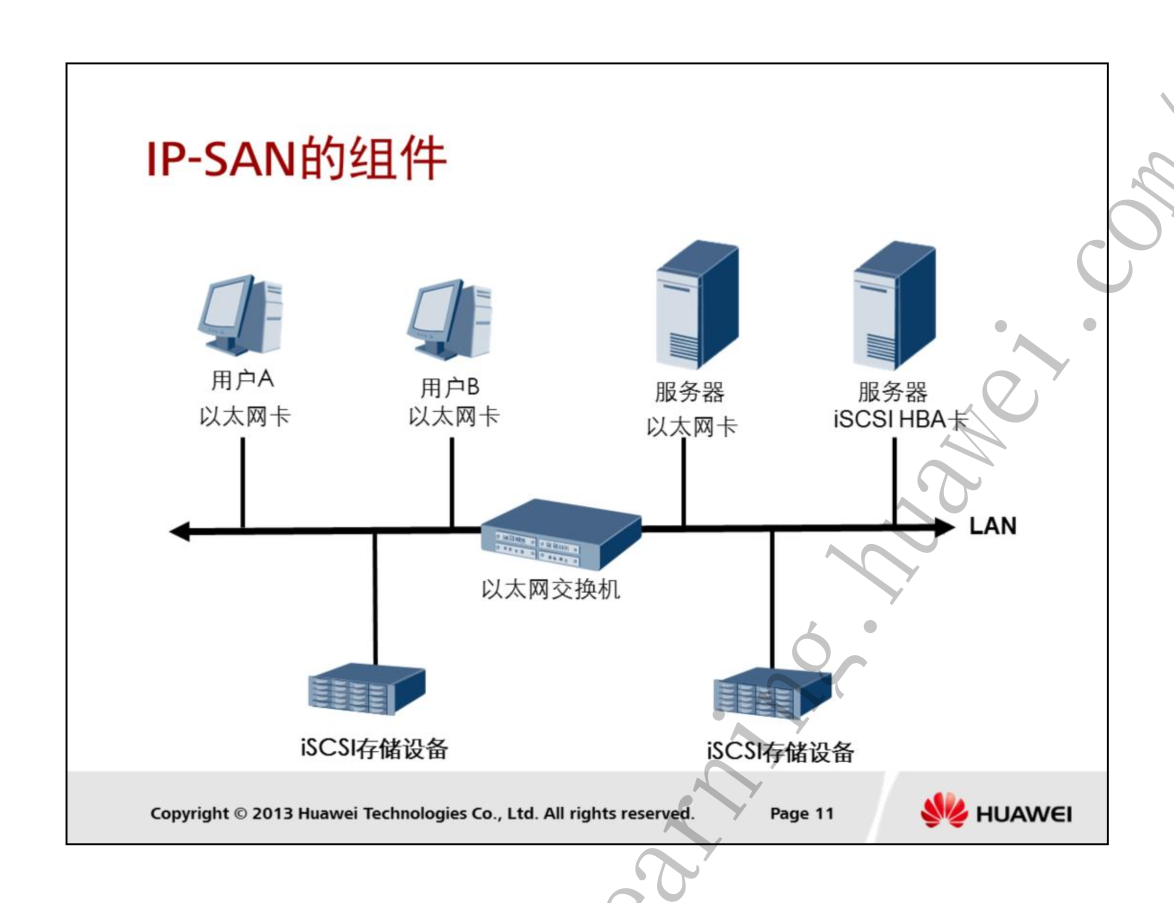


目录





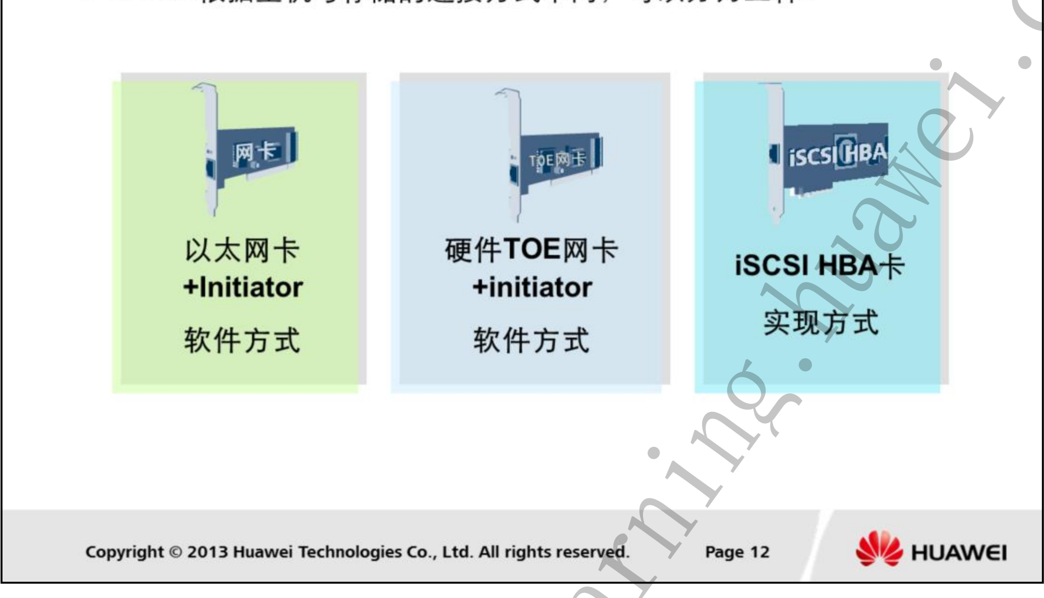
- 直连：
  - 主机与存储之间直接通过以太网卡、TOE卡或iSCSI HBA卡连接，这种组网方式简单、经济，但较多的主机分享存储资源比较困难；
- 单交换：
  - 主机与存储之间由一台以太网交换机，同时主机安装以太网卡或TOE卡或iSCSI HBA卡实现连接。这种组网结构使多台主机能共同分享同一台存储设备，扩展性强，但交换机处存在间点故障；
- 双交换：
  - 同一台主机到存储阵列端可由多条路径连接，扩展性强，避免了在以太网交换机处形成单点故障。



IP-SAN基于十分成熟的以太网技术，由于设置配置的技术简单、低成本的特色相当明显，而且普通服务器或PC机只需要具备网卡，即可共享和使用大容量的存储空间。 由于是基于IP协议的，所以能容纳所有IP协议网络中的部件。用户可以在任何需要的地方创建实际的SAN网络，而不需要专门的光纤通道网络在服务器和存储设备之间传送数据。同时，因为没有光纤通道对传输距离的限制，IP-SAN使用标准的TCP/IP协议，数据即可在以太网上进行传输。IP-SAN网络对于那些要求流量不太高的应用场合以及预算不充足的用户，是一个非常好的选择。

IP-SAN的组成部分：

- iSCSI存储设备
- 以太网交换机
- 以太网卡和iscsi initiator软件
- 以太网网线

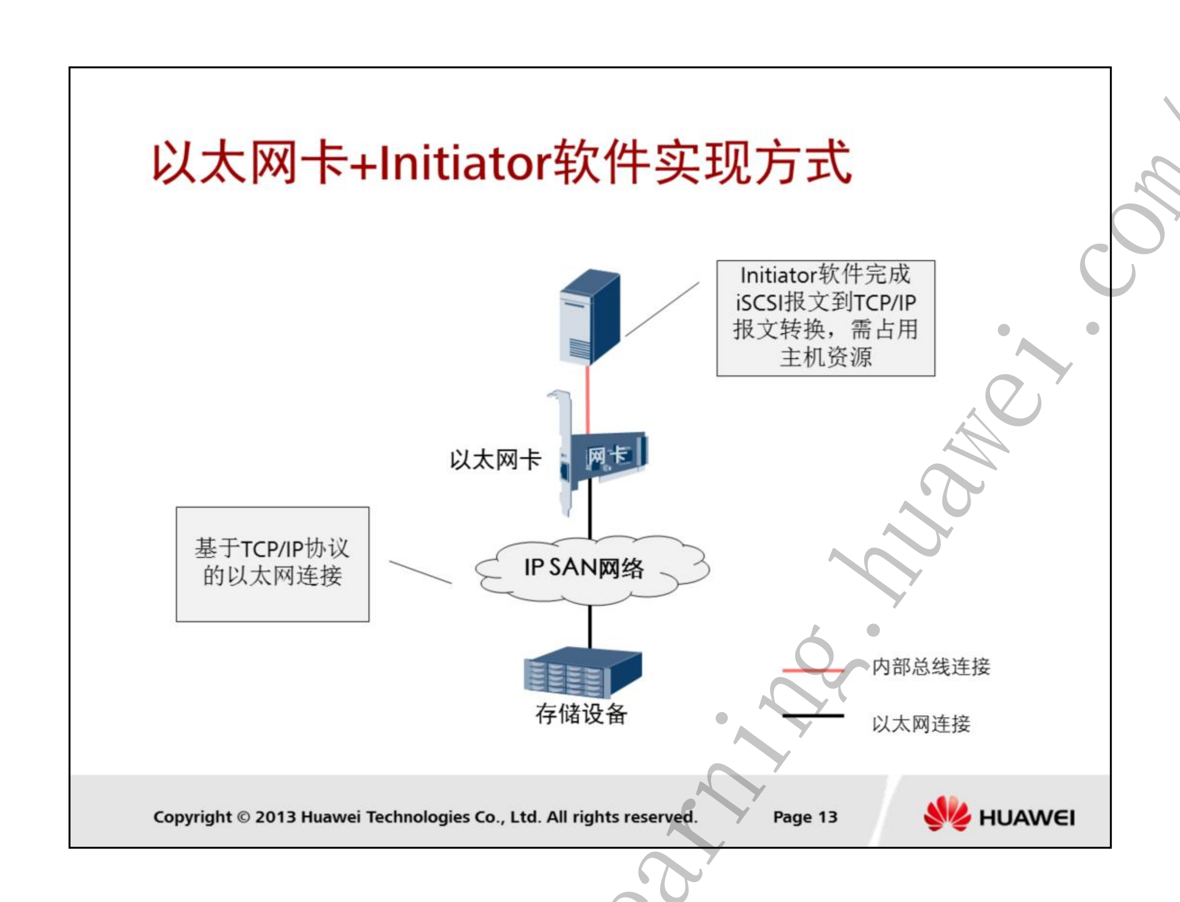


iSCSI设备通常使用IP接口作为其主机接口，并可以通过与传统以太网交换机的连接，构建一个基于TCP/IP协议的存储区域网络。根据主机端所采用的连接方式的不同，iSCSI设备与主机的连接通常有三种形式：

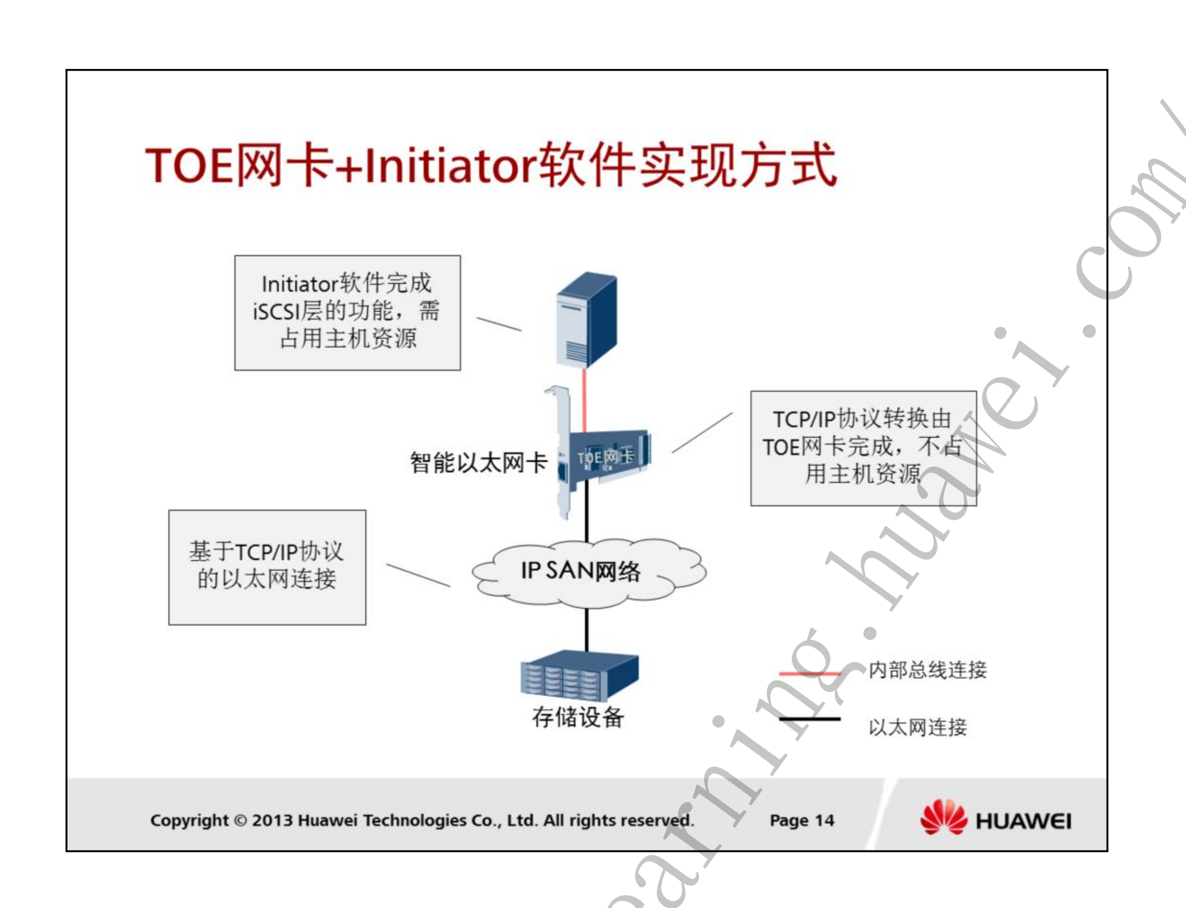
以太网卡＋Initiator软件方式：采用这种方式的主机使用标准的以太网卡（NIC）与网络进行连接。iSCSI层和TCP/IP协议栈功能通过主机CPU运行软件计算完成。由于这种方式直接使用传统主机系统通用的NIC卡，所以成本最低，但是由于需要占用CPU资源进行iSCSI协议和TCP/IP协议处理，所以会导致主机系统性能的下降。

TOE＋Initiator软件方式：采用这种方式的主机使用TOE（TCP offload Engine，TCP卸载引擎）网卡，iSCSI协议的功能仍然由主机的CPU完成，但是TCP协议处理则交由TOE网卡完成，从而有效减轻了主机端的负担。

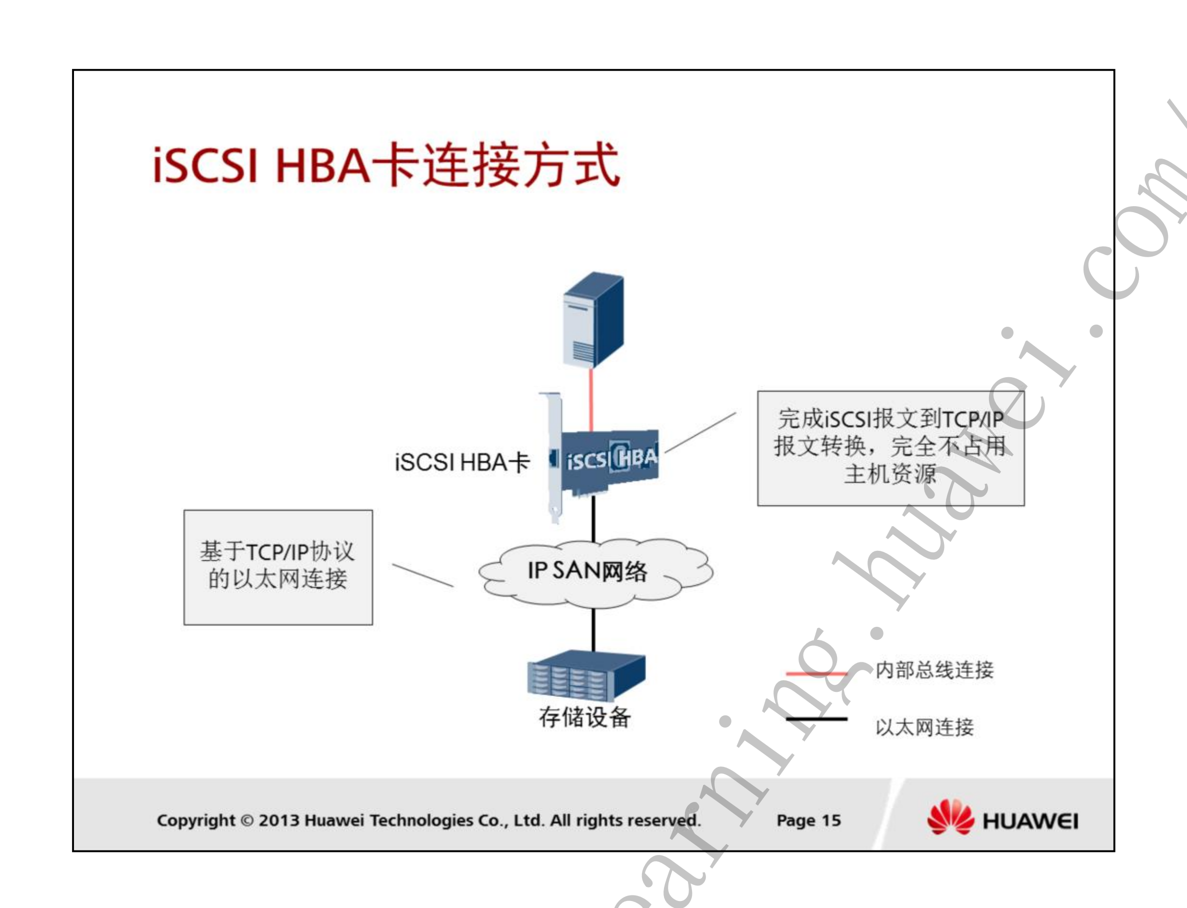
iSCSI HBA卡方式：采用这种方式的主机，其iSCSI协议功能及TCP/IP协议栈功能均由iSCSI HBA卡完成，对主机的开销占用最小。



服务器、工作站等主机设备使用标准的以太网卡，通过以太网线直接与以太网交换机连接，iSCSI存储也通过以太网线连接到以太网交换机上，或直接连接到主机的以太网卡上。在主机上安装Initiator软件以便将以太网卡虚拟为iSCSI卡，用以接收和发送iSCSI数据报文，从而实现主机和iSCSI设备之间的iSCSI协议和TCP/IP协议传输功能。由于采用普通的标准以太网卡和以太网交换机，无需额外配置适配器，因此此种方式硬件成本最低。缺点是进行iSCSI包文和TCP/IP包文转换需要占用主机端的资源，使主机的运行开销增加而导致系统性能下降。不过在对于I/O和带宽性能要求较低的应用环境中基本能够满足数据访问要求。



智能以太网卡可以将网络数据流量的处理工作全部转到网卡上的集成硬件中进行，TCP/IP协议栈功能由TOE网卡完成，而iSCSI层的功能仍旧由主机来完成，由此，采用TOE网卡可以大幅度提高数据的传输速率。与纯软件的方式相比较而言，这种方式部分降低了主机系统的运行开销而又不会使网络构建成本过多增加，是一种比较折衷的配置方案。



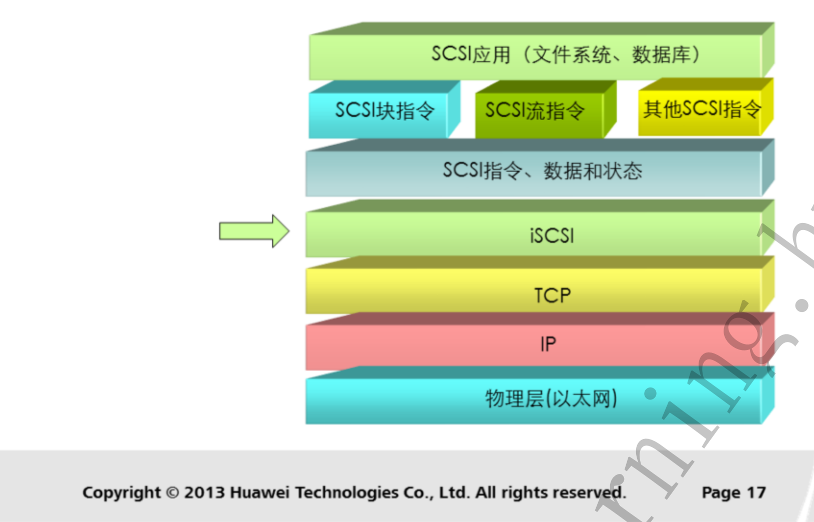
在主机上安装iSCSI HBA适配卡，从而实现主机与交换机之间、主机与存储设备之间的高效数据交换。iSCSI层和TCP/IP协议栈的功能均由主机总线适配器（HBA）来完成，对主机CPU的占用最少。这种方式数据传输性能最好，但是系统构建价格也最高。



目录

1. IP SAN产生与发展
2. IP SAN组成和组网连接
3. **iSCSI协议介绍**
4. FC协议与TCP协议融合
5. 华为IP SAN存储的实现与应用





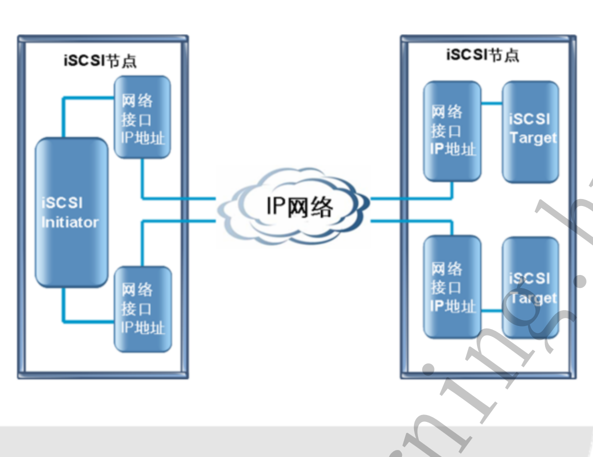
通过SCSI控制卡的使用可以连接多个设备，形成自己的“网络”，但是这个“网络”仅局限于与所附加的主机进行通信，并不能在以太网上共享。那么，如果能够通过SCSI协议组成网络，并且能够直接挂载到以太网上，作为网络节点和其它设备进行互联共享，那么SCSI就可以得到更为广泛的应用。所以，经过对SCSI的改进，就推出了iSCSI这个协议。基于iSCSI协议的IP-SAN是把用户的请求转换成SCSI代码，并将数据封装进IP包内在以太网中进行传输。

iSCSI方案最早是由Cisco和IBM两家发起，并且由Adaptec、Cisco、HP、IBM、Quantum等公司共同倡导。它提供基于TCP传输，将数据驻留与SCSI设备的方法。iSCSI标准草案在2001年推出，并经过多次论证和修改，于2002年提交IETF，在2003年2月，iSCSI标准正式发布。iSCSI技术的重要贡献在于其对传统技术的继承和发展：其一，SCSI（Small Computer System Interface，小型计算机系统接口）技术是被磁盘、磁带等设备广泛采用的存储标准，从1986年诞生起到现在仍然保持着良好的发展势头；其二，沿用TCP/IP协议，TCP/IP在网络方面是最通用、最成熟的协议，且IP网络的基础建设非常完善。这两点为iSCSI的无限扩展提供了坚实的基础。

IP网络的普及性将使得数据可以通过LAN、WAN或者是通过Internet利用新型IP存储协议传输，iSCSI既是在这个思想的指导下进行研究和开发的。iSCSI是基于IP协议的技术标准，实现了SCSI和TCP/IP协议的融合，对众多的以太网用户而言，只需要极少的投资，就可以方便、快捷地对信息和数据进行交互式传输和管理。

iSCSI体系结构

- iSCSI节点将SCSI指令和数据封装成iSCSI包，然后该数据封装被传送给TCP/IP层，再由TCP/IP协议将iSCSI包封装成IP协议数据以适合在网络中传输。



 Page 18



在支持iSCSI的系统中，用户在一台SCSI存储设备上发出存数据或取数据的命令，操作系统对该请求进行处理，并将该请求转换成一条或者多条SCSI指令，然后再传给目标SCSI控制卡。iSCSI节点将指令和数据封装（Encapsulation）起来，形成一个iSCSI包，然后该数据封装被传送给TCP/IP层，再由TCP/IP协议将iSCSI包封装成IP协议数据以适合在网络中传输。也可以对封装的SCSI命令进行加密处理，然后在不安全的网络上传送。

数据包可以在局域网或Internet上传送。在接收存储控制器上，数据报重新被组合，然后存储控制器读取iSCSI包中的SCSI控制命令和数据发送到相应的磁盘驱动器上，磁盘驱动器再执行初始计算机或应用所需求的功能。如果发送的是数据请求，那么将数据从磁盘驱动器中取出进行封装后发送给发出请求的计算机，而这整个过程对于用户来说都是透明的。尽管SCSI命令的执行和数据准备可以通过使用标准TCP/IP和现成的网络控制卡的软件来完成，但是在利用软件完成封装和解封装的情况下，在主机处理器上实现这些功能需要很多的CPU运算周期来处理数据和SCSI命令。如果将这些事务交给专门的设备处理，则可以将对系统性能的影响减少到最小程度，因此，发展在iSCSI标准下并执行SCSI命令和完成数据准备的专用iSCSI适配器是有必要的。iSCSI 适配器结合了NIC和HBA的功能。这种适配器以块方式取得数据，利用 TCP/IP处理引擎在适配卡上完成数据分化和处理，然后通过IP网络送出IP数据包。这些功能的完成使用户可以在不降低服务器性能的基础上创建一个基于IP的SAN。

iSCSI的发起端与目标端

- 发起端（Initiator）
  - SCSI层负责生成CDB（命令描述符块），将CDB传给iSCSI
  - iSCSI层负责生成iSCSI PDU（协议数据单元），并通过IP网络将PDU发给target
- 目标器（Target）
  - iSCSI层收到PDU，将CDB传给SCSI层
  - SCSI层负责解释CDB的意义，必要时发送响应

iSCSI
Initiator | Target
SCSI | SCSI
iSCSI | iSCSI
TCP | TCP
IP | IP
Link | Link

 Page 19 HUAWEI

iSCSI的通信体系仍然继承了SCSI的部分特性，在iSCSI通信中，具有一个发起I/O请求的启动器设备（Initiator）和响应请求并执行实际I/O操作的目标器设备（Target）。在Initiator和Target建立连接后，Target在操作中作为主设备控制整个工作过程。

- iSCSI Initiator

iSCSI启动器，可分为三种，即软件Initiator驱动程序、硬件的TOE（TCP Offload Engine，TCP卸载引擎）卡以及iSCSI HBA卡。就性能而言，软件Initiator驱动程序最差、TOE卡居中、iSCSI HBA卡最佳。

- iSCSI Target

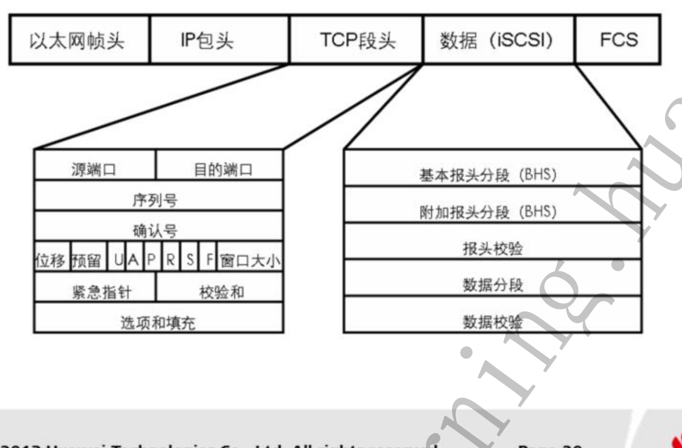
iSCSI目标器iSCSI Target，通常为iSCSI 磁盘阵列、iSCSI 磁带库等。

iSCSI协议为Initiator和Target定义了一套命名和寻址方法。所有的iSCSI节点都是通过其iSCSI名称被标识的。这种命名方式使得iSCSI名称不会与主机名混淆。

iSCSI使用iSCSI Name来唯一鉴别启动设备和目标设备。地址会随着启动设备和目标设备的移动而改变，但是名字始终是不变的。建立连接时，启动设备发出一个请求，目标设备接收到请求后，确认启动设备发起的请求中所携带的iSCSI Name是否与目标设备绑定的iSCSI Name一致，如果一致，便建立通信连接。每个iSCSI节点只允许有一个iSCSI Name，一个iSCSI Name可以被用来建立一个启动设备到多个目标设备的连接，多个iSCSI Name可以被用来建立一个目标设备到多个启动设备的连接。

iSCSI数据包封装模型

- 所有的SCSI命令都被封装成iSCSI协议数据单元
- iSCSI利用TCP/IP协议栈中传输层的TCP协议为连接提供可靠的传输机制



 Page 20



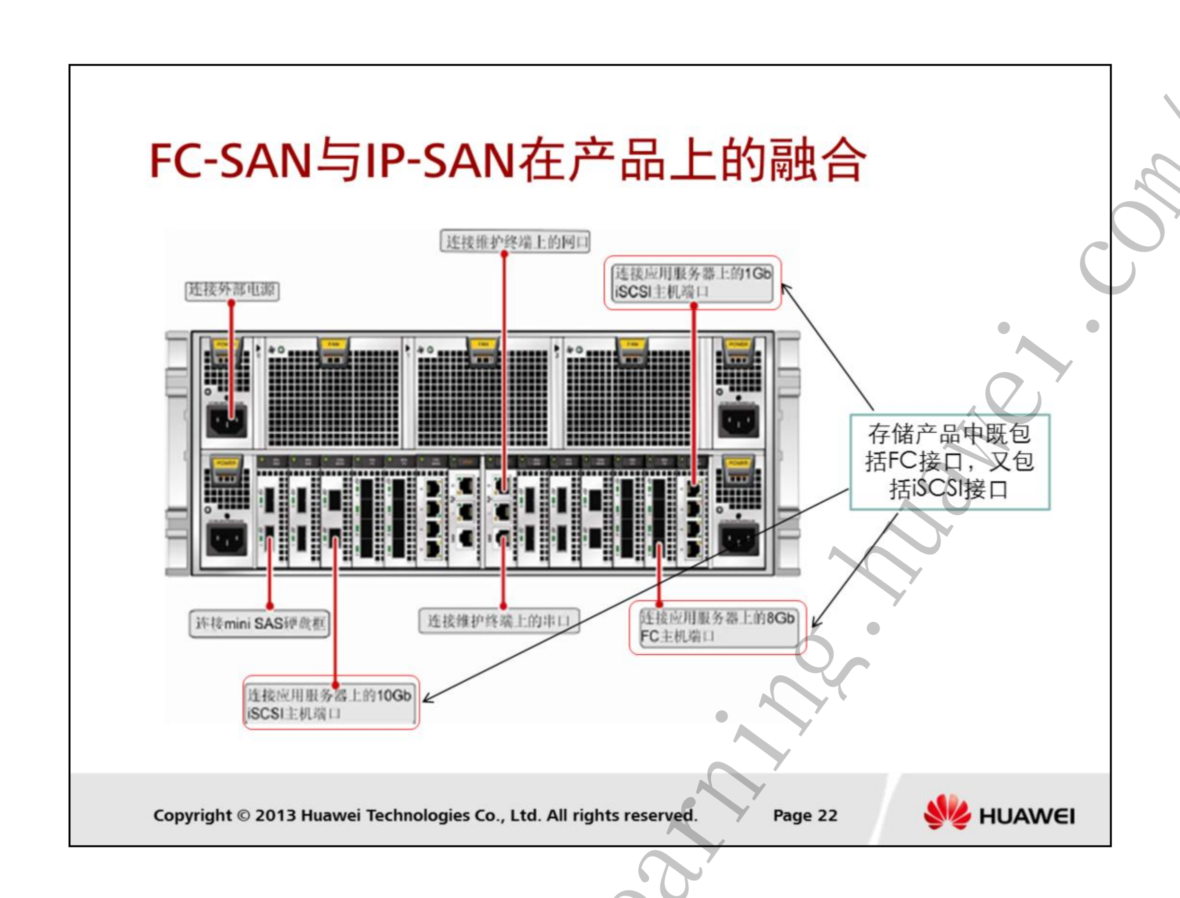
支持iSCSI的服务器可以配置一块专用的iSCSI主机总线适配器卡。所有的SCSI命令都被封装成iSCSI协议数据单元（Protocol Data Unit，PDU），iSCSI会利用TCP/IP协议栈中传输层的TCP协议为连接提供可靠的传输机制，再封装TCP数据段头以及IP数据包头后，其内部所封装的SCSI命令或数据对于底层网络设备而言是不可见的，网络设备只会将其视为普通IP数据包进行传递，从而实现了SCSI指令和数据的透明传输。

iSCSI协议是将SCSI的远程过程调用（Remote Procedure Call，RPC）映射到IP协议的过程。iSCSI协议提供了独立于其所携带的SCSI CDB的层的概念。iSCSI请求传递SCSI命令，iSCSI响应处理SCSI响应和状态。iSCSI为基于IP协议的PDU提供了一个在SCSI的命令结构内映射的机制，SCSI的命令及参数被填充在一定长度的数据块内进行传输。一个iSCSI翻译器取得SCSI CDB（Command description Block，命令描述块），并将其映射为iSCSI PDU，在TCP连接上发送到一个目标iSCSI设备。翻译器通过连接ID识别一组映射SCSI连接的TCP连接。从启动设备和目标设备的角度来看，这个连接就像是一个普通的SCSI通信一样，整个IP传输对于启动器设备和目标设备而言是透明的。启动设备或目标设备可以是一个iSCSI设备，能够用TCP直接在IP网中通信。

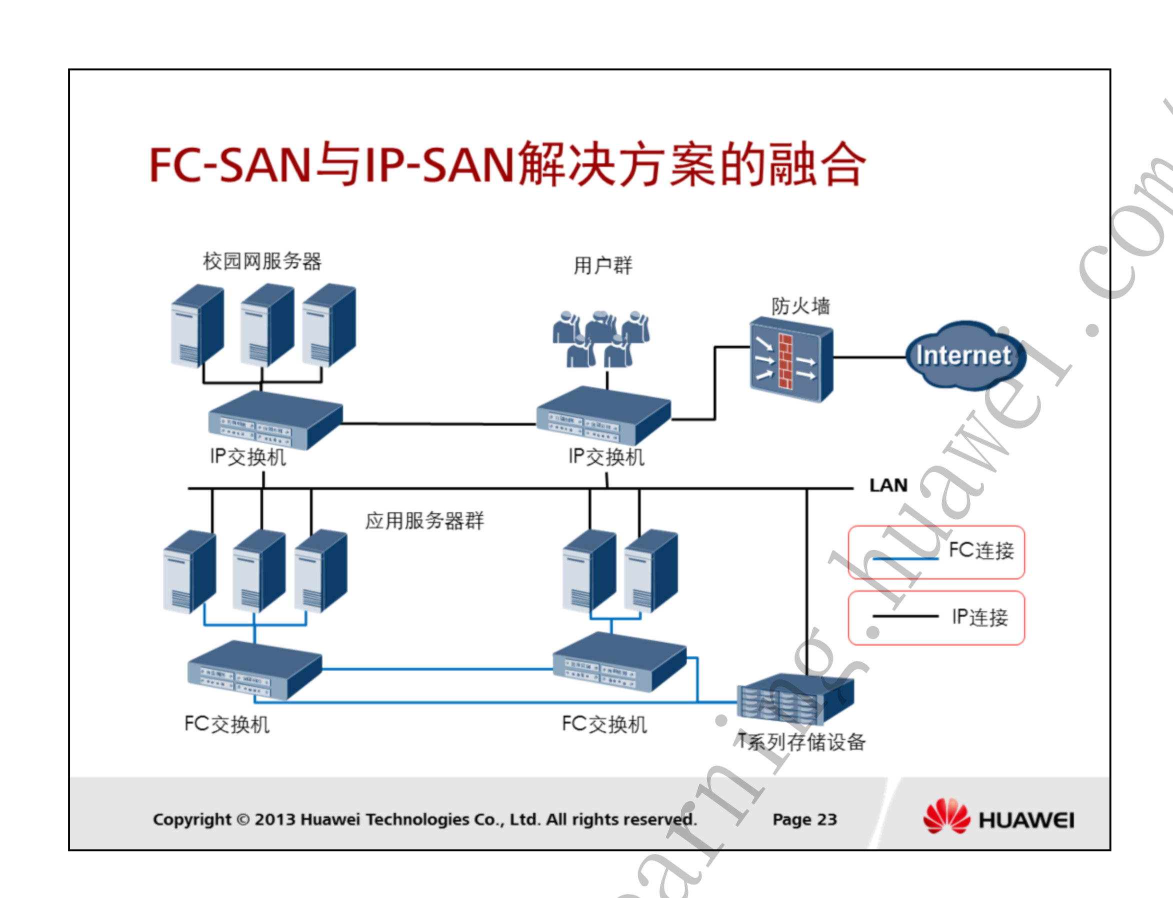


目录





如磁盘阵列上控制器上即包含FC端口、又包含iSCSI口，能同时满足FC-SAN的组网需求，也能满足IP-SAN的组网需求，这是不完全的融合。



统一存储解决方案典型T系列产品，可以同时对外提供FC-SAN、IP-SAN和NAS共享。

FC与TCP协议融合

- 目前FC与TCP/IP协议的真正融合主要有两种趋势：
  - TCP/IP网络承载FC信道
    - FCIP
    - iFCP
    - FCOE
  - 以FC信道承载TCP/IP数据
    - IPFC
- 从现有的情况来看，以太网技术和FC技术都在飞速发展，IP-SAN和FC-SAN会在很长的一段时间内都将是并存且互为补充的。

 Page 24 

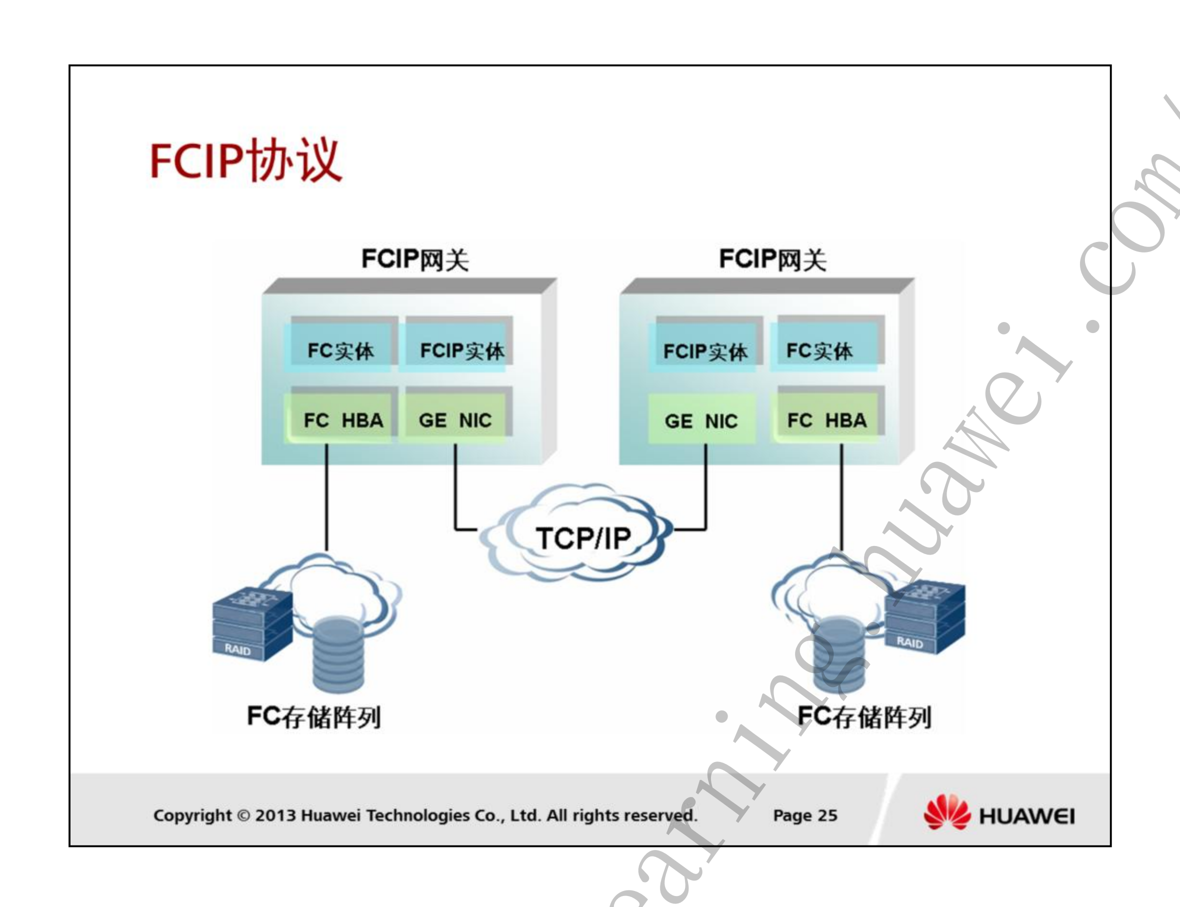
FCIP(Fibre Channel over IP)，基于IP的光纤通道（FCIP）是连接TCP/IP网络链路上的光纤通道架构的一项IETF建议标准。利用目前的IP协议和设施来连接两个异地FC SAN的隧道，用以解决两个FC SAN的互连问题。

iFCP(Internet Fibre Channel Protocol)，Internet光纤信道协议（iFCP）是一种网关到网关的协议，为TCP/IP网络上的光纤设备提供光纤信道通信服务。iFCP 使用 TCP 提供拥塞控制、差错监测与恢复功能。iFCP主要目标是使现有的光纤信道设备能够在IP网络上以线速互联与组网。此协议及其定义的帧地址转换方法允许通过透明网关（transparent gateway）将光纤信道存储设备附加到基于 IP 的网络结构。

FCoE(Fibre Channel over Ethernet)，它利用以太网路，传送光纤通道（Fibre Channel）的信号，让光纤通信的资料可以在10 Gigabit以太网路网络骨干中传输，但仍然是使用光纤通道的协定。

IPFC(IP over Fiber Channel，在光纤通道上的IP)。IPFC使用在两个服务器之间的光纤通道连接作为IP数据交换的媒介。为此，IPFC定义了如何通过光纤通道网络传送IP分组。跟所有的应用协议一样，IPFC实现为在操作系统中的一个设备驱动器。对本地IP配置的连接使用“ifconfig”或“ipconfig”。然后IPFC驱动器寻址光纤通道主机适配卡，就可以在光纤通道上发送IP分组。

下面我们着重介绍一下FCIP和iFCP协议。



FCIP（Fiber Channel over IP，基于IP协议的光纤通道）是基于IP协议传输的光纤通道数据帧的解决方案，是由Brocade、Gadzoox、Lucent、McData以及Qlogic公司共同提出的。FCIP这一技术的核心，是把光纤通道协议的数据帧封装在IP数据包里，以便在覆盖广阔的TCP/IP网络中进行传递。网络中的其它设备接收后，由专门的目标设备进行解封装还原FC数据帧。FCIP协议实质上就是采用隧道技术的IP-SAN方案。采用FCIP技术可以实现利用目前的IP网络来连接两个异地的FC-SAN，以解决FC-SAN之间的互联问题。这一隧道传输技术是通过使用FCIP网关来实现的，通过光纤通道交换机的扩展端口连接到每个FC-SAN上，所有发往远程站点的存储数据均通过公用的IP隧道。接收端的光纤通道交换机负责将到来的每个帧交换至目的光纤通道端点设备。

FCIP协议是一种隧道（tunnel）协议，IP地址和TCP连接只用在位于IP网络重点的FCIP网关设备上。FCIP能够为两个FC-SAN之间提供IP连接，但是不能为两个独立的FC存储设备提供IP连接，即FCIP不能通过IP协议实现FC存储设备端到端的连接。

FCIP协议利用IP网络中创建的“隧道”在两个FC-SAN网络之间实现FC协议的数据传输，将真正意义上的远程数据镜像和FC-SAN的灵活性以及IP网络的低成本和易用性结合在一起，降低了远程操作的成本和操作的复杂性。FCIP提供了在TCP/IP协议中封装FC协议数据帧的方法，消除了FC目前存在的距离限制，允许通过IP网络来互联FC-SAN，使得数据的访问变得更加灵活，存储策略的部署更加容易。

FCIP的协议栈

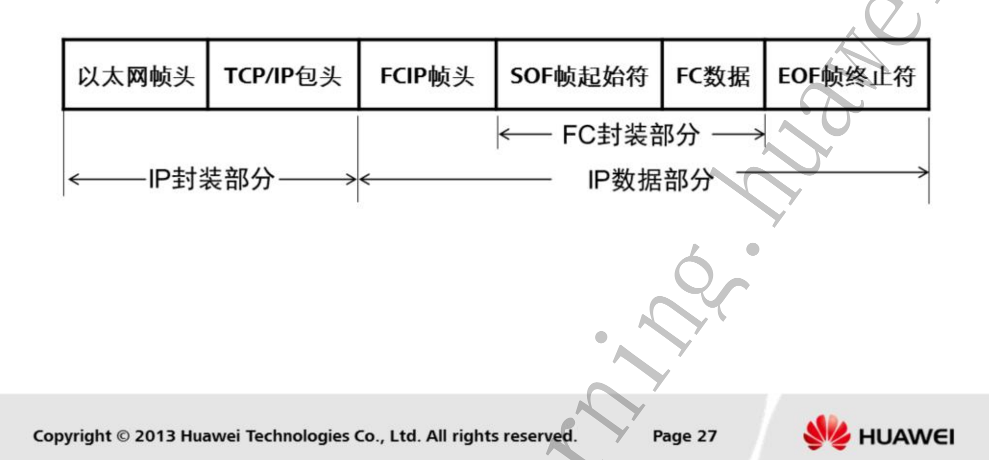
- 在FCIP的协议栈中，FCIP协议处于FC和TCP之间，也就意味着FCIP可以互联FC和TCP这两种协议网络。

SCSI 应用（文件系统、数据库）		
SCSI 块指令	SCSI 流指令	其它 SCSI 指令
SCSI 指令、数据和状态		
FCP		
FC		
FCIP		
TCP		
IP		
以太网		

 Page 26

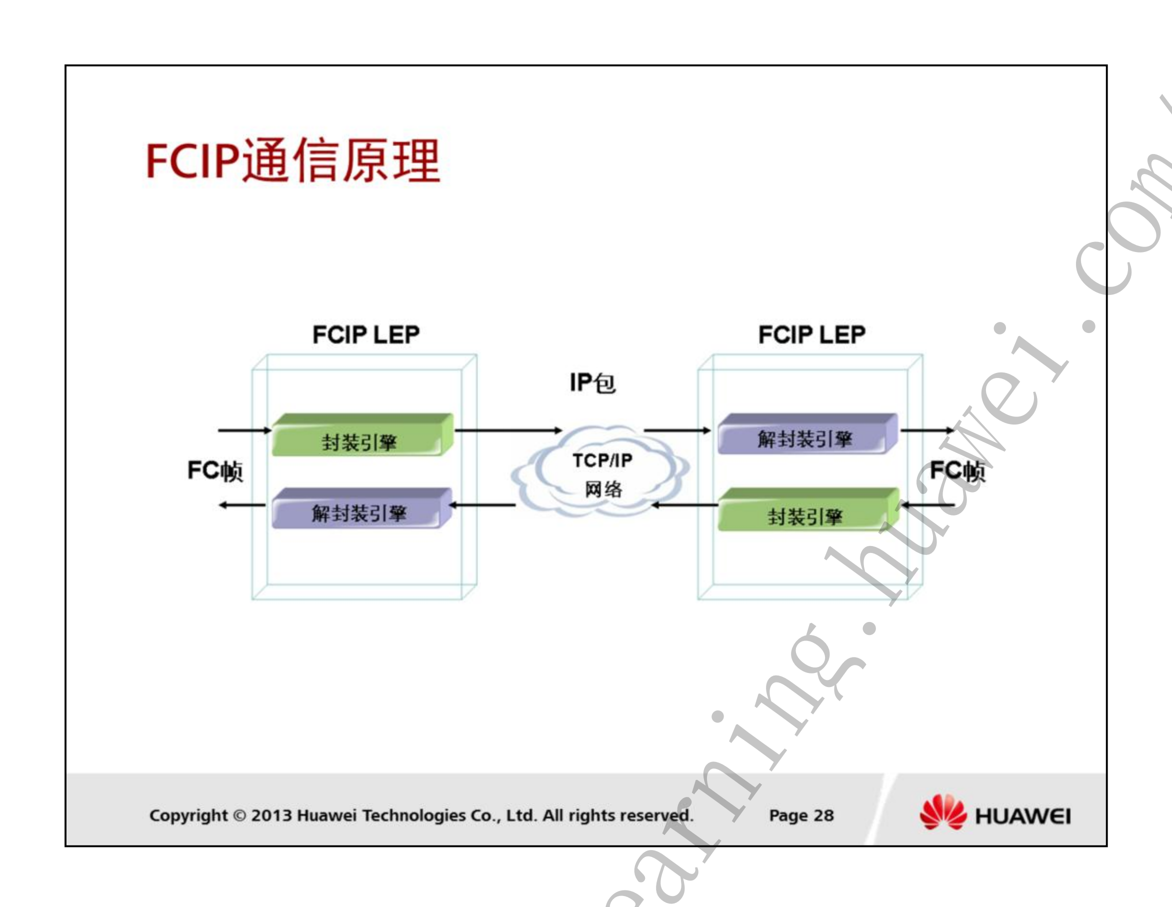


FCIP协议是一个点到点的隧道封装协议，它可以实现多个本地FC-SAN网络经由FCIP网关通过IP网络进行互联并对其进行管理。在FCIP的协议栈中，FCIP协议处于FC和TCP之间，也就意味着FCIP可以互联FC和TCP这两种协议网络。在TCP下层是IP协议和下层的数据链路以及物理层协议，而FC协议的上层则有FCP和SCSI协议，由此可见FCIP协议联系了底层的IP网络和高层的SCSI应用，实现了不同网络、不同协议之间的网络设备互联和应用的融合。



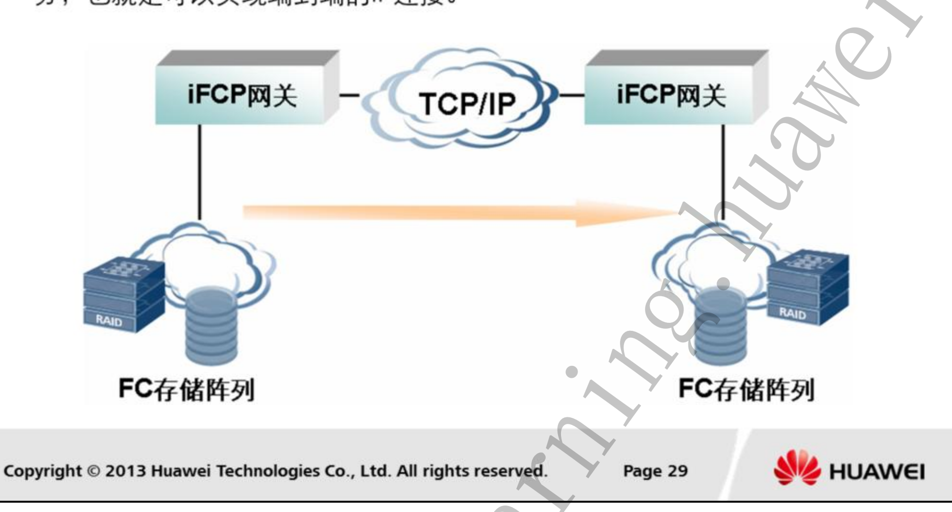
在FCIP数据封装中，光纤通道网络体系结构提供的终端寻址、地址解析、信息路由等均保持不变，而IP协议在这里只是作为传输协议用以承载FC数据帧在IP网络中进行传输。

FCIP在FC帧和TCP包头之间加入了FCIP包头、用来显示FCIP协议的版本、帧长度等字段。发送端FCIP网关设备将FC封装为FCIP帧，通过IP网络传送。接收端FCIP网关设备接收到FCIP帧后，解封装IP和TCP报头，还原成FC帧并通过一个或多个FC交换机发送到目的节点。



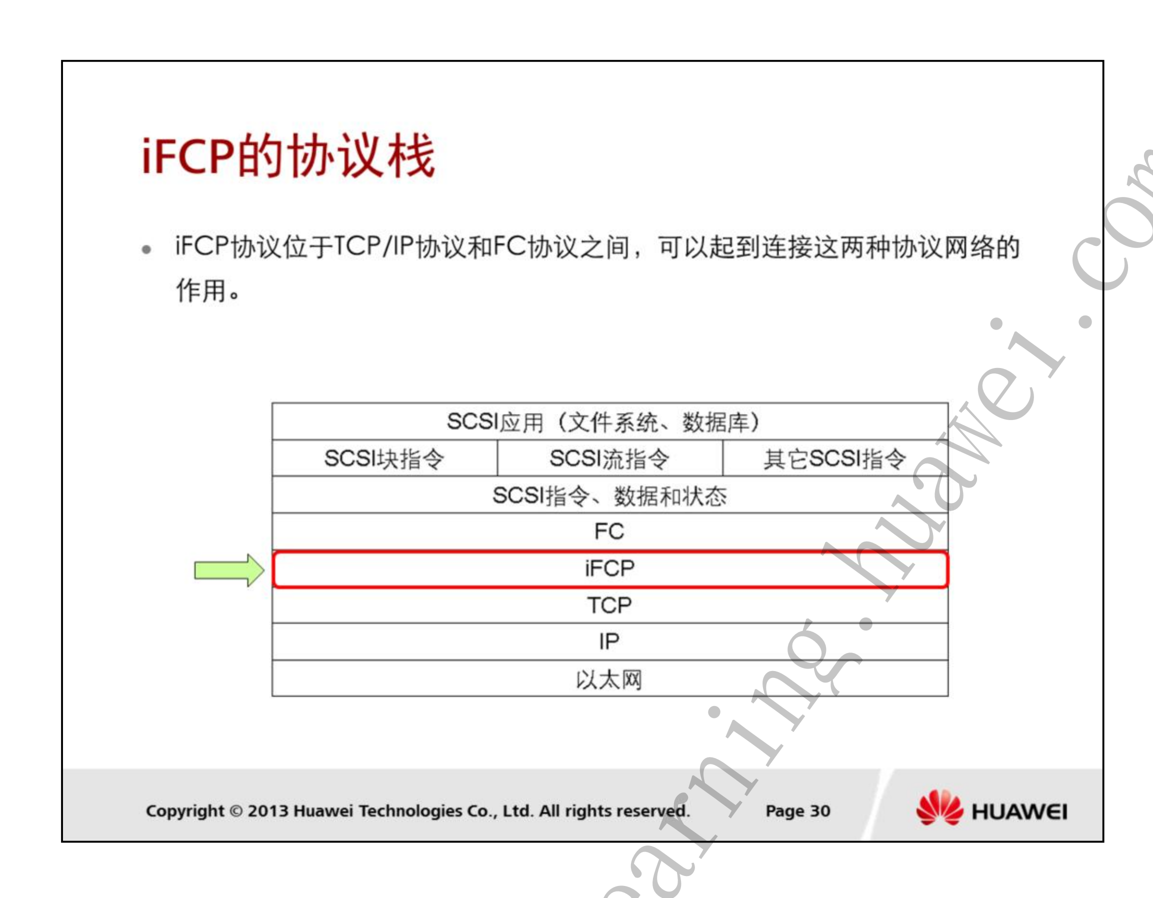
整个FCIP的通信过程是由其数据引擎推动进行的。首先在源FCIP连接端点（FCIP Link EndPoint，FCIP LEP）处对FC协议数据帧进行封装，然后通过TCP/IP协议在IP网络中进行传输，到达目的FCIP连接端点后进行解封装，读出其中的数据并执行其中的FC指令。

FCIP作为一种隧道技术，仍然存在一些缺点。首先，其带宽相对FC而言，由于利用的是IP通道，所以带宽仍远低于FC。其次，由于FC协议帧被封装进IP数据包中，但是IP网络智能管理工具并不能识别这些数据，使得很多很好的管理控制机制无法应用于FCIP，比如目录服务、流量控制和QoS等。最后，由于FCIP仅仅是在TCP/IP网络中构建起一个传输FC协议帧的隧道将两个远端的FC-SAN连接起来，它并没有解决单个FC-SAN的设备互操作性问题和管理问题，本地的SAN仍然是采用的FC技术。



iFCP（Internet Fibre Channel Protocol，互联网光纤通道协议）是一种网关到网关的协议，为TCP/IP网络上的光纤通道设备提供光纤信道通信服务，也就是可以实现端到端的IP连接。FC存储阵列、HBA（主机总线适配器，Host Bus Adapter）、交换机等可以直接连接到iFCP网关上。iFCP使用TCP提供流量控制、错误检测和错误恢复功能。iFCP主要目标是使现有的光纤信道设备能够在IP网络上实现高速互联与组网。iFCP及其定义的帧地址转换方法允许通过透明网关将FC存储设备挂载到基于TCP/IP 协议的网络结构中。

iFCP可以直接替代FC架构，通过iFCP存储交换机可以直接连接FC的各个设备并进行分组，而不仅仅是简单地对FC-SAN进行远程连接，但是iFCP不支持独立的存储区域网络的合并操作（Merge），因此无法组成单一的逻辑SAN 。iFCP的优势在于在建立连接的同时还能够建立网关分区，可以将出现故障的区域隔离开来，并克服了点到点隧道的限制。并且iFCP提供FC设备端到端的连接，TCP连接的中断只会影响到一个通信对，而不会影响到其他通信，也不会将一个设备的错误带给其他设备。基于iFCP实现了SAN的路由故障隔离、安全及灵活管理，具有比FCIP更高的可靠性。



iFCP协议位于TCP/IP协议和FC协议之间，可以起到连接这两种协议网络的作用。iFCP协议层的主要功能是在本地和远程N_PORT间传输光纤通道帧映像。当帧被传输到远程N_PORT时，iFCP层开始封装并路由光纤通道帧。光纤通道帧包括每一个光纤通道信息单元，通过预先建立的TCP连接在IP网络上传输。

iFCP协议封装

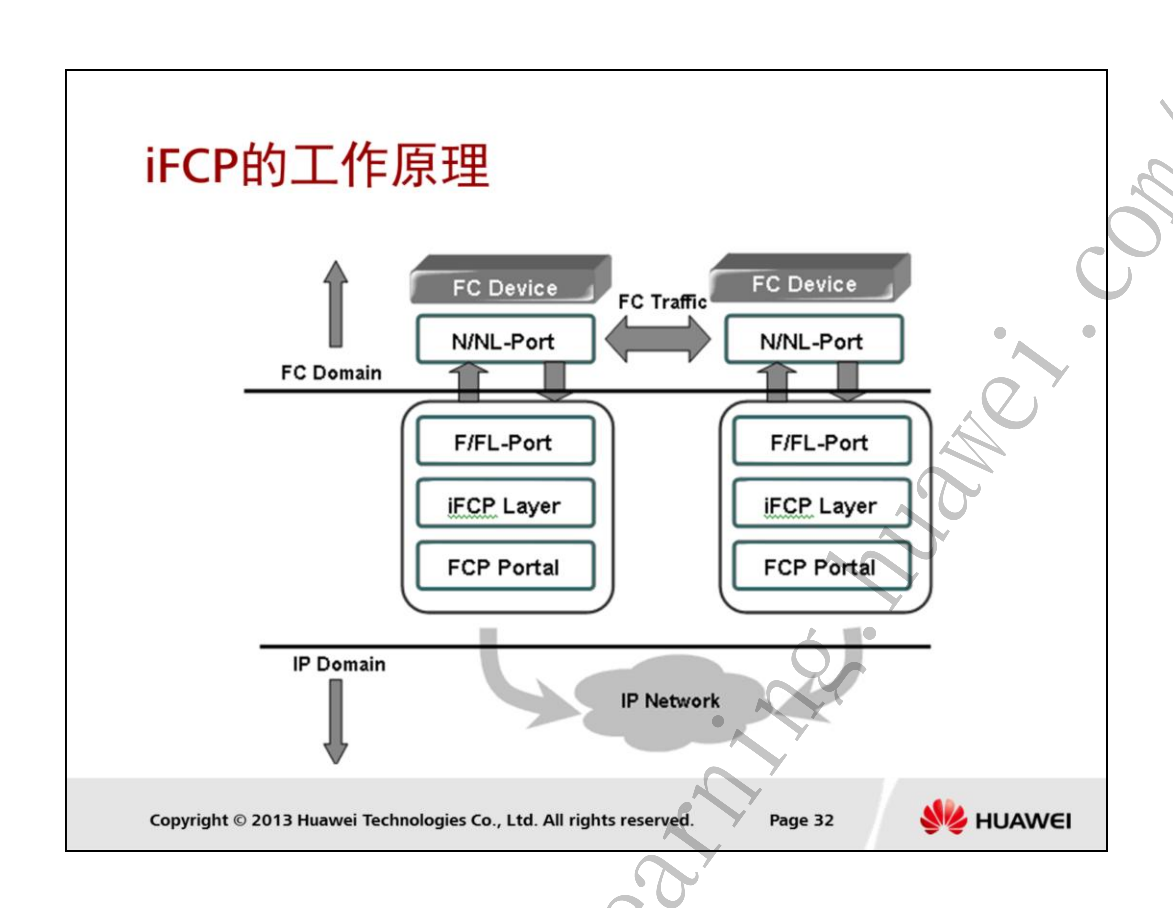
- 在iFCP层中，FC设备的24位fabric地址被映射到一个唯一的IP地址上，为Fibre Channel启动器和目标提供了本地IP地址的编址工作。

IP报头	TCP报头	iFCP报头	Fiber Channel帧

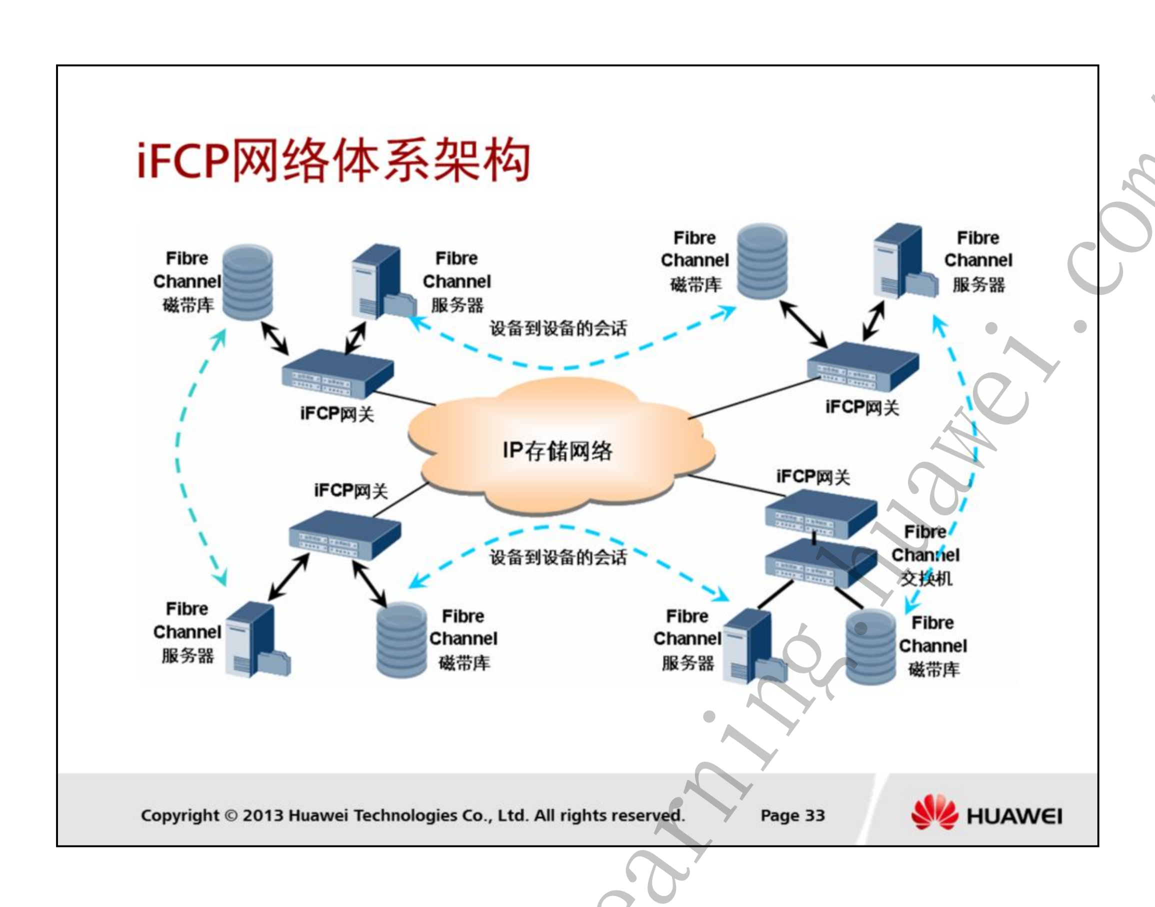
D_IP:01XXXX ⟷ D_ID:01XXXX
S_IP:02XXXX ⟷ S_ID:02XXXX

 Page 31 HUAWEI

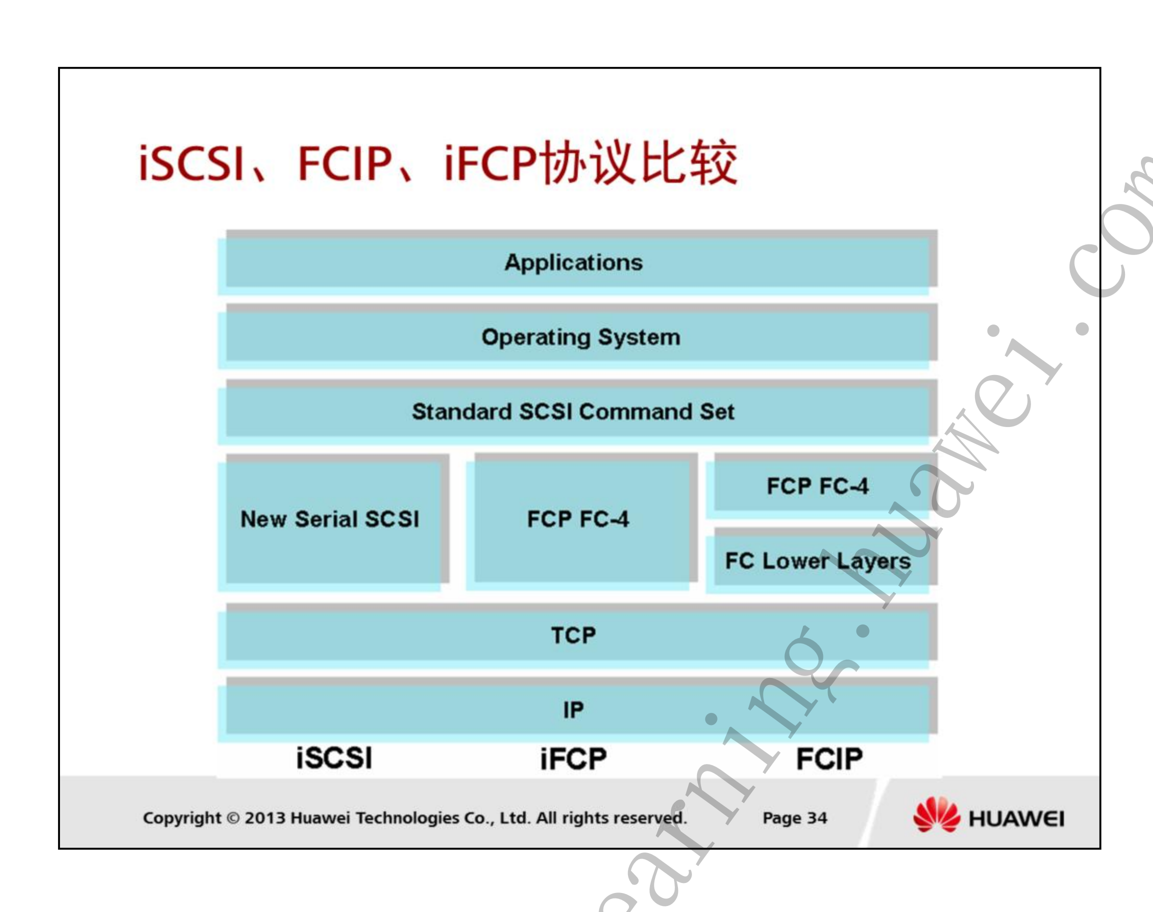
iFCP在FC帧和TCP包头之间，在iFCP层中，FC设备的24位fabric地址被映射到一个唯一的IP地址上，为Fibre Channel启动器和目标提供了本地IP地址的编址工作。iFCP代替了Fibre Channel的底层传输层（FC-2），它使用TCP/IP在IP网络上进行可靠传输。



iFCP的工作原理是，将Fibre Channel数据以IP包形式封装，并将IP地址映射到分离光纤通道设备。由于在IP网中每类光纤通道设备都有其独特标识，因而能够与位于IP网其他节点的设备单独进行存储数据收发。光纤通道信号在iFCP网关处终止，信号转换后存储通信在IP网中进行，这样iFCP就打破了传统光纤通道网的距离（在不使用中继器的情况下，FC的传输距离约为10公里）限制。



在使用iFCP协议构建的IP-SAN存储网络中，存储设备没有被限制在光纤通道SAN的IP网络中分布。iFCP存储交换机直接接替FC-SAN中的光纤通道交换机，这就意味着iFCP交换机也具有SNS（存储名称服务器）功能，为终端节点提供名称发现服务。在iFCP交换机中指派4字节的IP地址给每一个光纤通道终端节点。当光纤通道设备发送一个SNS名称查询时，这个请求将被iFCP交换机截住，并由iSNS服务器进行解释。在光纤通道层，一个适用的目标地址表将返回给发起者，此时其余IP的光纤通道地址表就映射光纤通道地址，以便相应的IP地址可以通过IP网络传到目标设备。



FCIP和iSCSI技术在IP存储领域是两个相对的极端，FCIP可以看作是FC的扩展策略，它包含了部分的IP协议信息和大量的FC成分，所以从严格意义上来说FCIP并不能算真正的IP存储策略。而iSCSI协议目的则是要用IP协议完全取代FC协议在SAN中的应用，iSCSI协议中完全不含有FC的内容，只包含了IP信息。iSCSI与iFCP相比较具有一定的相似性，iSCSI和iFCP都在存储端设备中采用了IP协议技术，而不同在于，iSCSI为通过IP网络进行块数据传输，定义了其本身的串行SCSI的实现。这三种协议都位于TCP/IP和SCSI协议之间，为TCP/IP和SCSI的沟通起到了纽带的作用。

FCoE协议

- 直接在增强型无损以太网基础设施上传输光纤信道信号功能的协议。
- FCoE把FC帧封装在以太网帧中，允许LAN和SAN的业务流量在同一个以太网中传送。

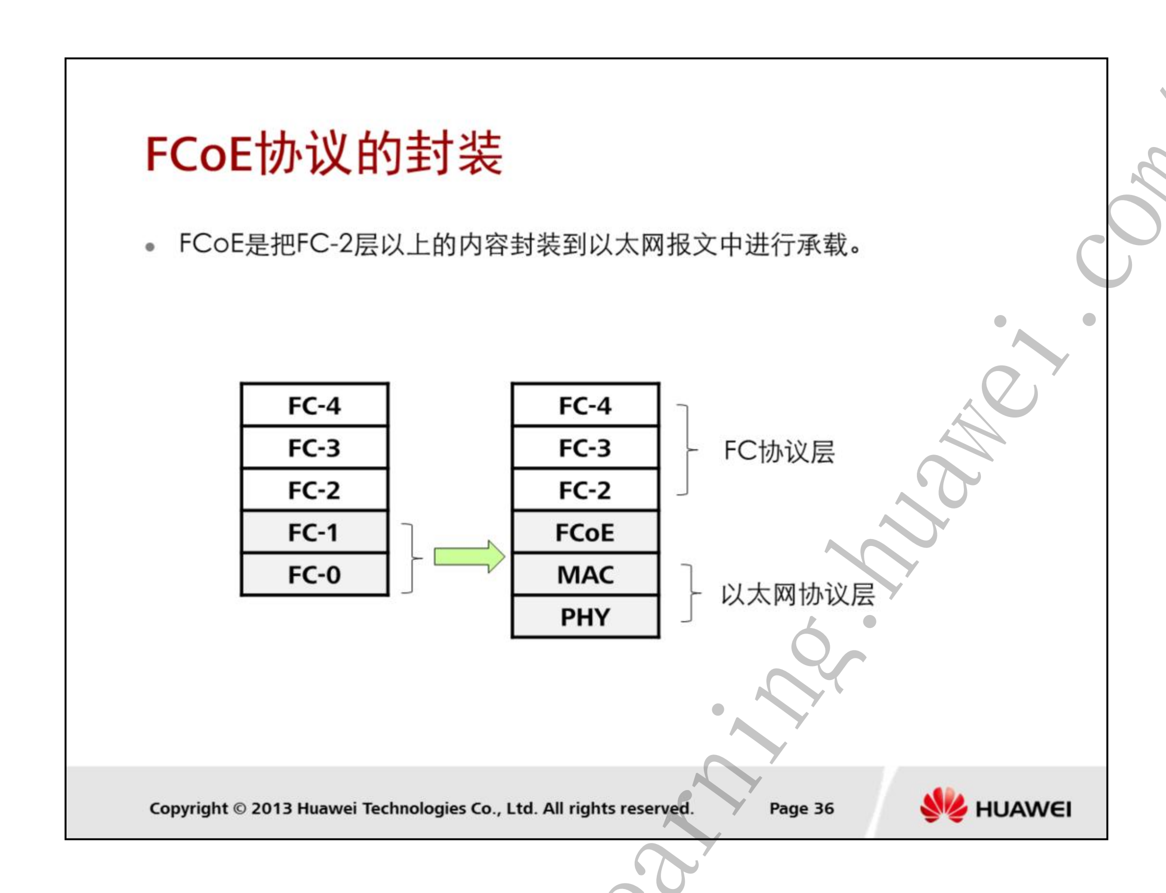
以太网数据链路层帧

- 业务流IP
- 块存储FCoE
- 网络电话VoIP
- 视频流VoIP

 Page 35 HUAWEI

FCoE（Fibre Channel over Ethernet）可以提供标准的光纤通道原有服务，如发现、全局名称命名、分区等，而且这些服务都可以照标准原有的运作，保有FC原有的低延迟性、高性能。

从FC协议的角度来看，FCoE就是把FC承载在一种新型的链路上，即以太网二层链路；从以太网的角度来看，FCoE仅是其承载的另外一种上层协议（类似于IP/IPX）。



FC网络协议栈有五层，其中FC0定义承载介质类型，FC1定义帧编解码方式，FC2定义分帧协议和流控机制，FC3定义通用服务，FC4定义上层协议到FC的映射。

融合增强型以太网（CEE）

- FCoE采用增强型以太网作为物理网络传输架构，能够提供标准的光纤通道有效内容载荷。
- 融合增强型以太网（CEE）可以避免类似TCP/IP协议的开销和数据包损失。



Page 37



传统的以太网是一种尽力服务的网络模式，当网络拥塞时有可能发生丢包，进而导致出现数据包重传或超时现象。FCoE网络融合技术的出现，对以太网提出了无丢包服务的要求。为此，IEEE 802.1和IETF标准组织制定一些新的标准，创建一个新的、更强大的以太网协议系列，即融合增强型以太网（CEE）。

- 基于优先级的流量控制（PFC）

以太网Pause机制能够实现网络不丢包的要求，但它会阻止一条链路上的所有流量，PFC是对Pause机制的一种增强。PFC可以在一条以太物理链路上创建8个独立的虚拟链路，并允许单独暂停和重启其中任意一条虚拟链路。这一方法使网络能够为单个虚拟链路创建无丢包类别的服务，使其能够与同一接口上的其他类型的流量共存。

- 增强的传输选择（ETS）

ETS可以为为不同的业务流量设定优先级和保证带宽，并允许低优先级的流量使用高优先级队列闲置的带宽，这样可以提高整个网络的效率。

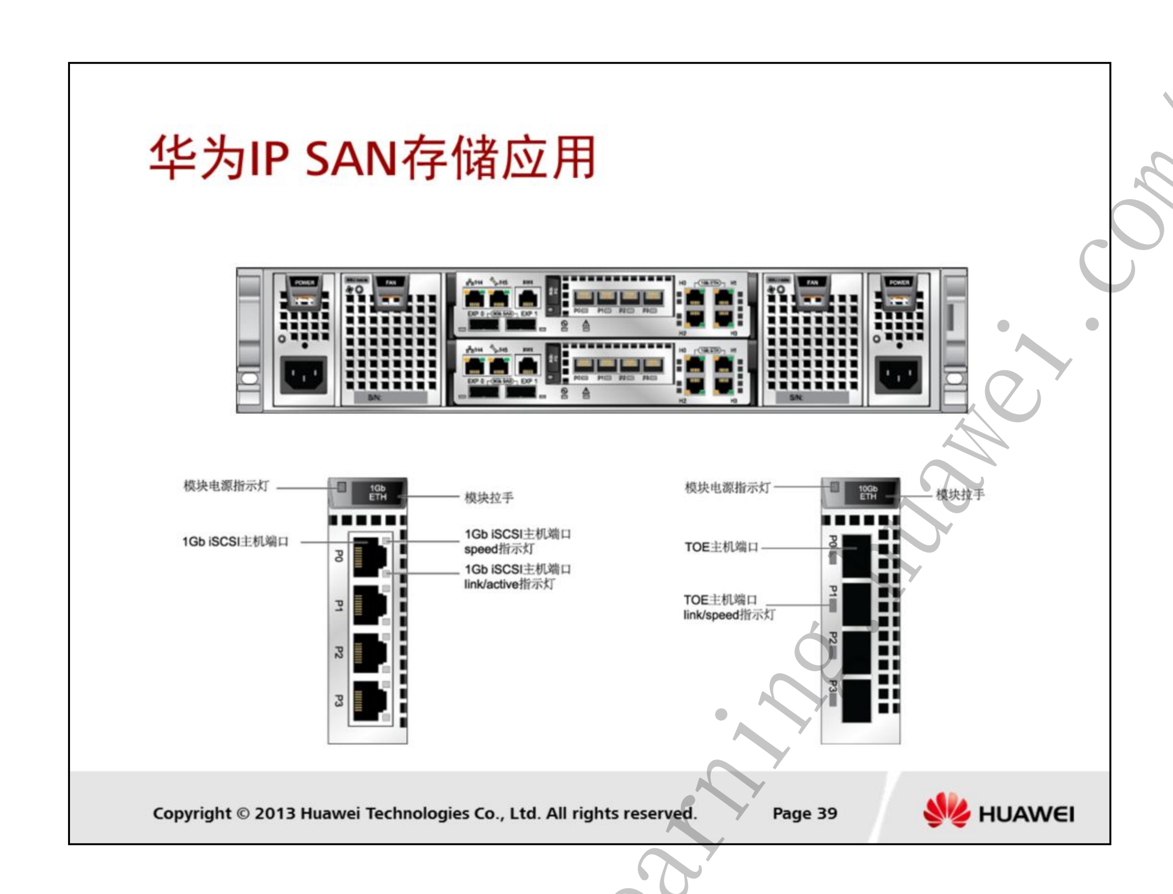
- 拥塞通告

当网络中发生拥塞时，由拥塞点向数据源发送指示来限制引起拥塞的流量，并在拥塞消失时通知其取消限制。拥塞通知提供了一种在二层网络对持续拥塞的流量端到端管理方法。



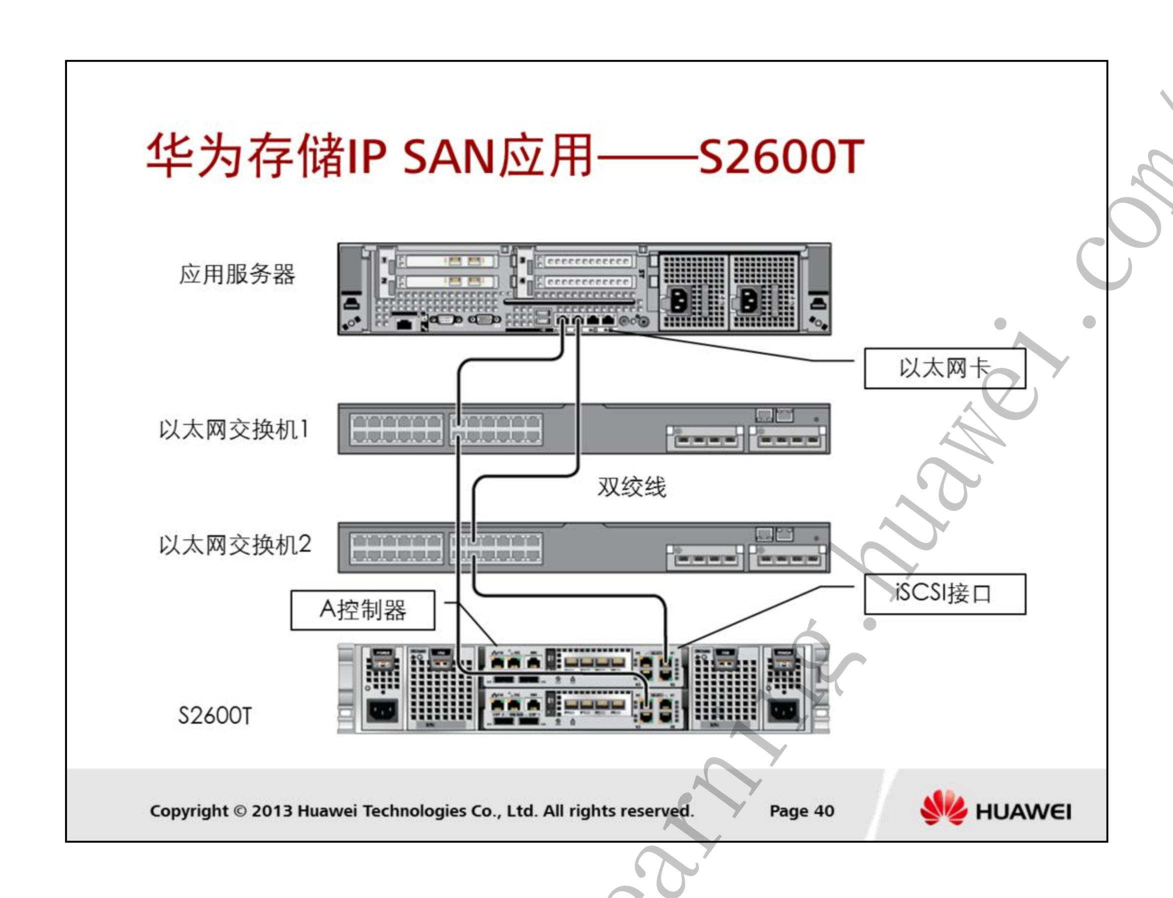
目录





1Gb iSCSI接口模块提供了应用服务器与存储系统的业务接口，用于接收应用服务器发出的数据读写指令。1Gb iSCSI接口模块提供4个传输速率为1Gbit/s的iSCSI接口，用于接收应用服务器发出的数据交换命令。

10Gb TOE接口模块提供了应用服务器与存储系统的业务接口，用于接收应用服务器发出的数据读写指令。10Gb TOE接口模块提供4个传输速率为10Gbit/s的TOE接口，用于接收应用服务器发出的数据交换命令。



华为S2600T存储阵列通过网线或光纤将控制框iSCSI主机端口与应用服务器端口进行连接，建立业务通道，实现存储设备与应用服务器之间的数据交换。iSCSI主机端口有1Gbit/s和10Gbit/s两种速率，分别使用网线和光纤连接。

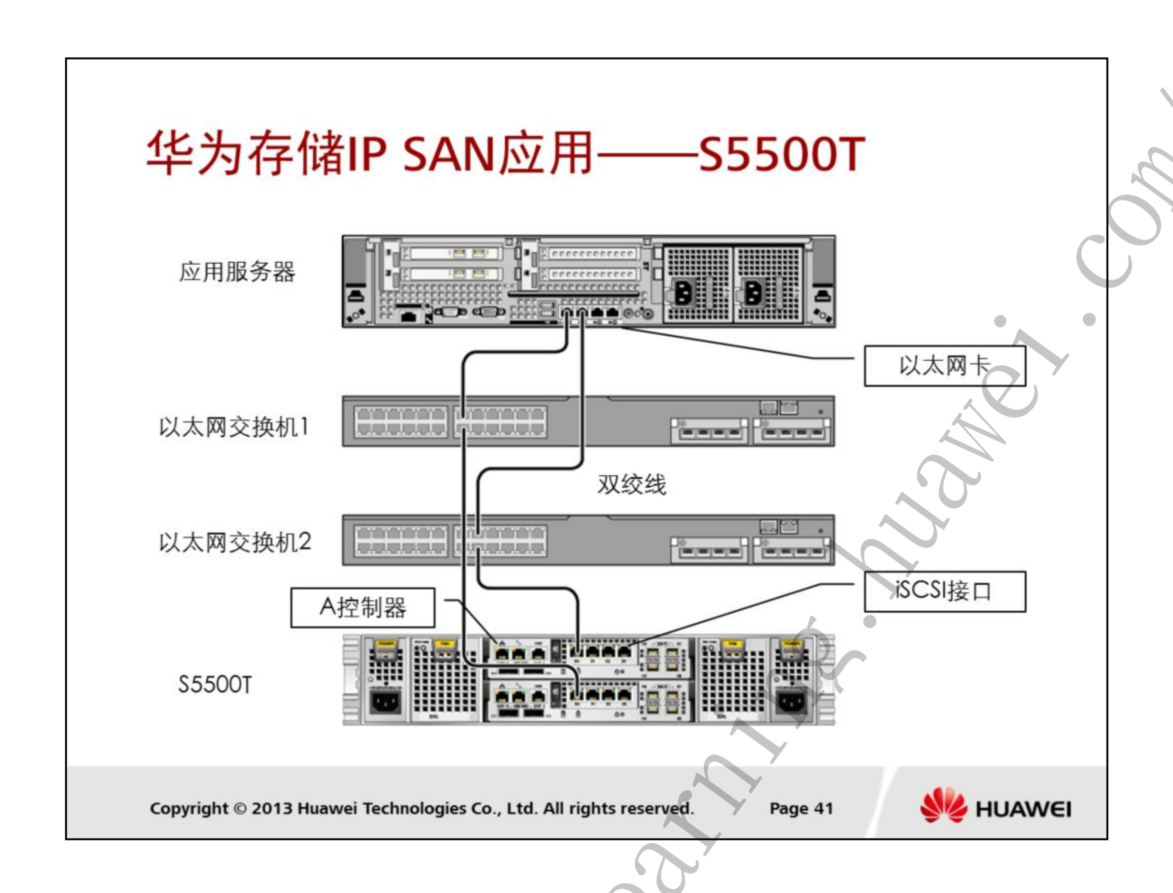
当1Gb iSCSI主机端口的速率设置为1000Mbit/s时，不支持半双工模式。

存储设备与应用服务器的连接方式有两种：

- 直接连接
- 通过以太网交换机连接

通过以太网交换机连接时，建议您提前做好以下规划：

- 交换机网段的规划，并确定使用网段对应的交换机端口。
- 交换机的首尾端口常用于交换机之间的组网，不建议用于主机与存储设备连接。



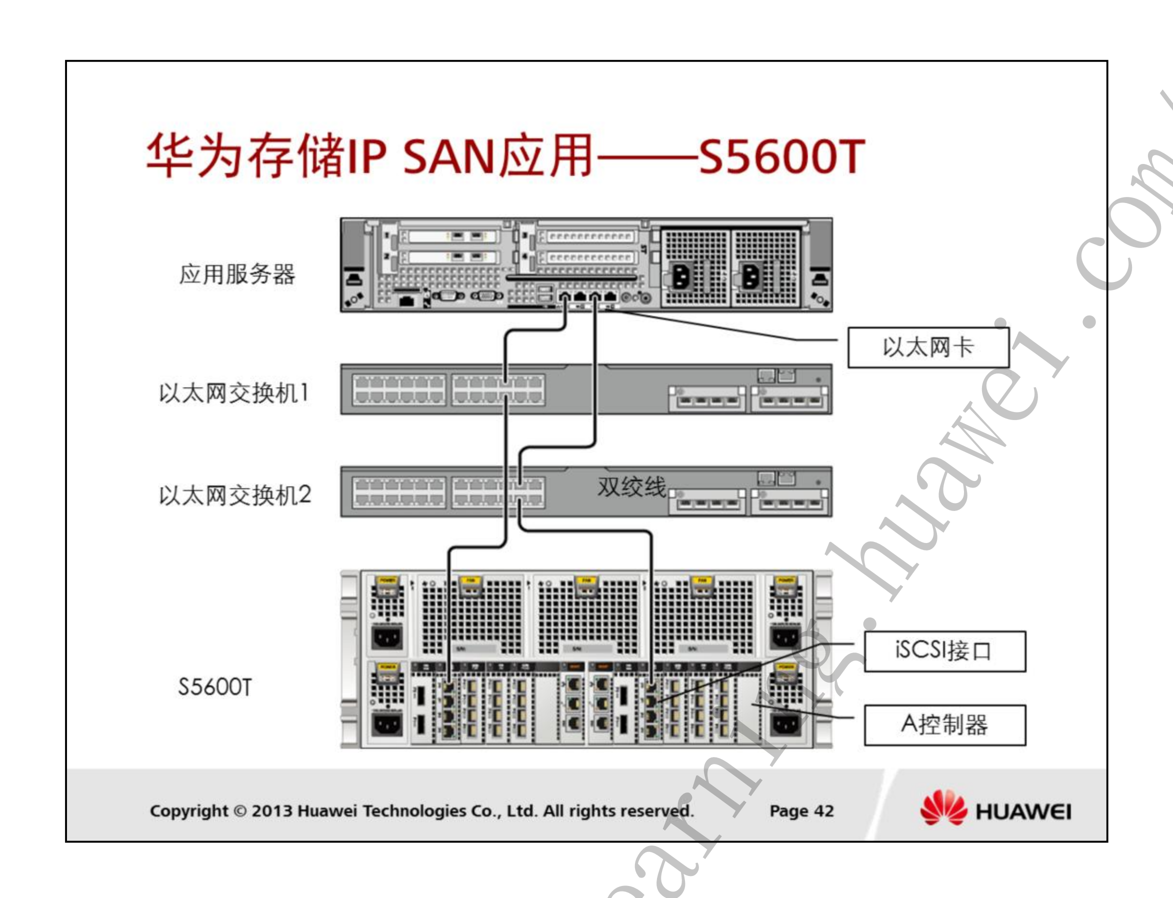
华为S5500T存储阵列通过网线或光纤将控制框iSCSI主机端口与应用服务器端口进行连接，建立业务通道，实现存储设备与应用服务器之间的数据交换。iSCSI主机端口有1Gbit/s和10Gbit/s两种速率，分别使用网线和光纤连接。

与S2600T类似，华为5500T存储阵列设备与主机服务器的连接方式有两种：

- 直接连接
- 通过以太网交换机连接

通过以太网交换机连接时，建议您提前做好以下规划：

- 交换机网段的规划，并确定使用网段对应的交换机端口。
- 交换机的首尾端口常用于交换机之间的组网，不建议用于主机与存储设备连接。



华为S5600T存储阵列设备控制框为4U盘控分离设计，通过网线或光纤将控制框iSCSI主机端口与应用服务器端口进行连接，建立业务通道，实现存储设备与应用服务器之间的数据交换。iSCSI主机端口有1Gbit/s和10Gbit/s两种速率，分别使用网线和光纤连接。

与S2600T类似，华为5600T存储阵列设备与主机服务器的连接方式有两种：

- 直接连接
- 通过以太网交换机连接

通过以太网交换机连接时，建议您提前做好以下规划：

- 交换机网段的规划，并确定使用网段对应的交换机端口。
- 交换机的首尾端口常用于交换机之间的组网，不建议用于主机与存储设备连接。

硬盘框级联和管理网口及串口管理组网在第6章中已经介绍，这里面不在重复。



总结

- IP SAN产生与发展
- IP SAN组成和组网连接
- iSCSI协议介绍
- FC协议与TCP协议融合
- 华为IP SAN存储的实现与应用





思考题

1. IP SAN与FC SAN的主要区别有哪些?
2. IP SAN主要有哪几部分组成?
3. IP SAN的连接方式分哪几种？各有什么特点?
4. iSCSI协议的发起端（Initiator）和目标器（Target）起到什么作用?
5. 描述FCIP、iFCP和FCoE协议的特点和应用。





习题

- 判断题

  1. 存储IP-SAN组网主机采用以太网卡+Initiator软件实现方式时，TCP/IP协议转换操作由以太网卡完成，不占用主机资源。 （T of F)

- 多选题

  1. 以下对FC SAN与IP SAN比较，说法正确的是：

     2. FC SAN受到光纤传输距离的限制，而IP SAN理论上是无限制的。

     B. FC SAN网络因为是隔离的，不与外网直接相连，故比较安全。

     C. FC SAN投资成本高，而IP SAN可以利用现有TCP/IP网络，投入相对低。

     D. FC SAN维护相对简单，而IP SAN基于TCP/IP其维护比较复杂。

 Page 45



习题答案：

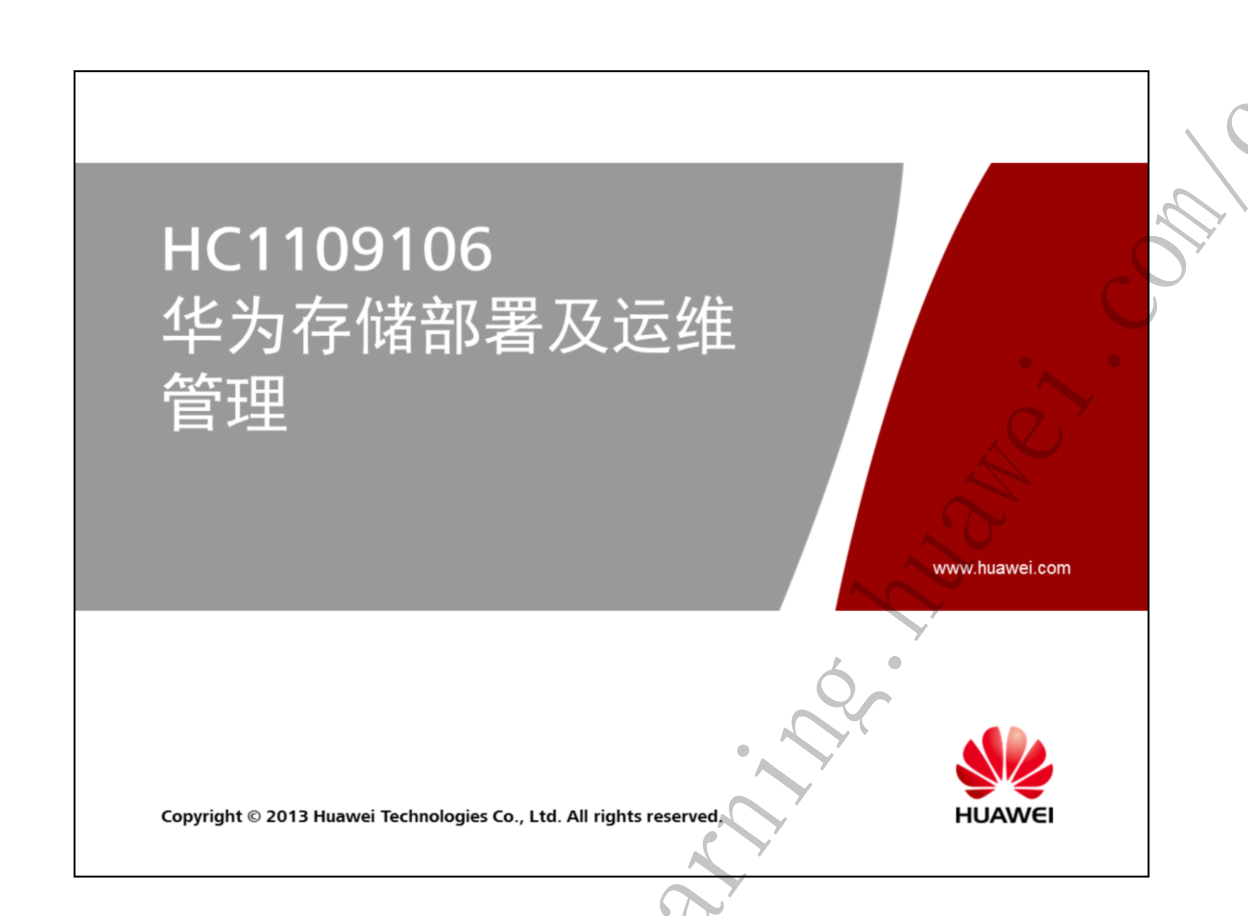
- 判断题：1.F
- 多选题：1.ABC

Thank you

www.huawei.com

HC1109106 华为存储部署及运维管理




HC1109106
华为存储部署及运维管理
www.huawei.com
Copyright © 2013 Huawei Technologies Co., Ltd. All rights reserved.
HUAWEI



目标

- 学习完本课程后，您将能够：
  - 掌握存储初始化配置流程
  - 掌握存储端基础配置
  - 掌握主机端基础配置
  - 掌握存储运维管理方法





目录

1. **存储初始化配置流程**
2. 存储端基础配置
3. 主机端基础配置
4. 存储运维管理



存储系统基础配置流程

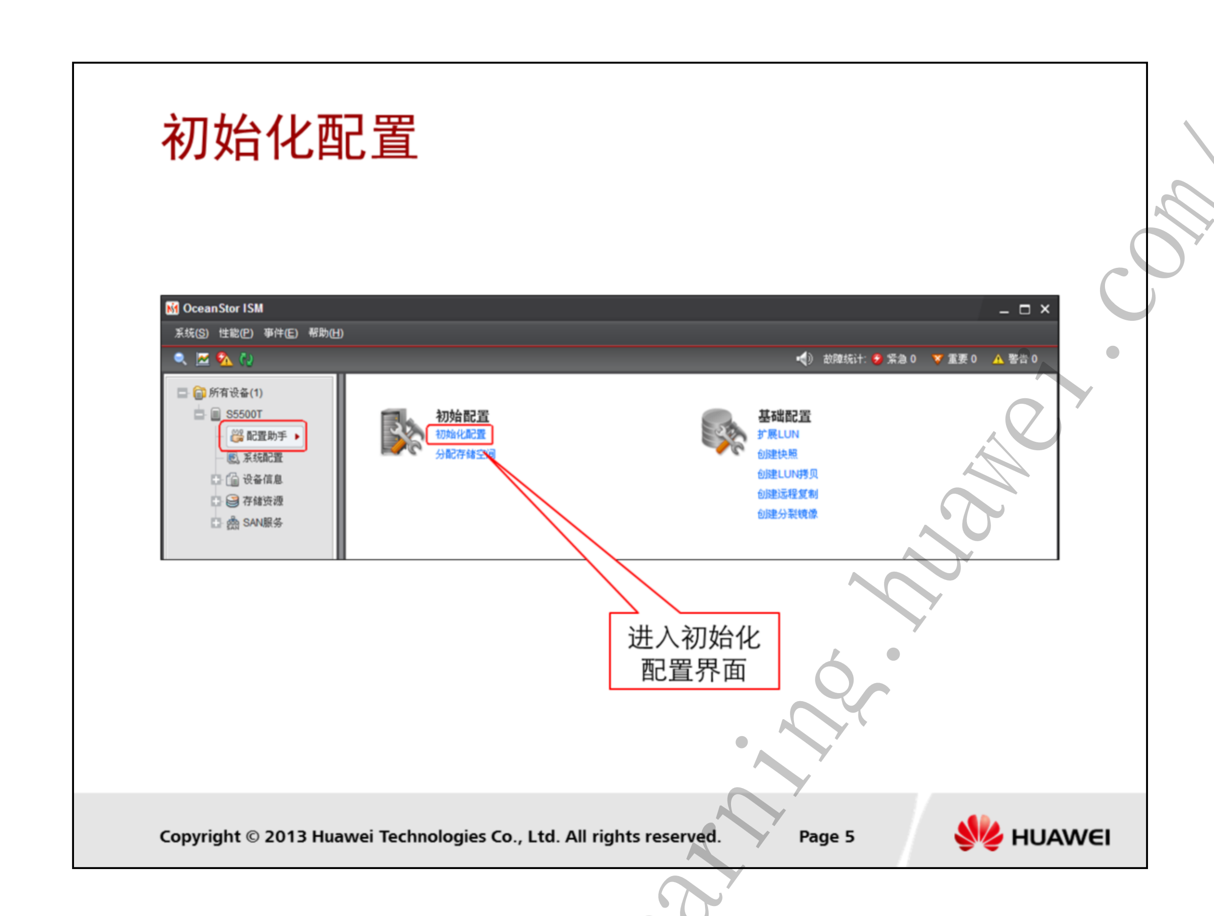
开始
↓
创建存储空间 → 创建RAID → 创建热备盘 → 创建LUN
↓
分配存储空间 → 创建主机组 → 创建主机 → 添加映射
↓
建立连接
- FC → 添加启动器
- iSCSI → 配置启动器 → 添加启动器

↓
使用存储空间
- Windows → 格式化 → 建立分区
- Linux → 格式化 → 建立分区 → mount

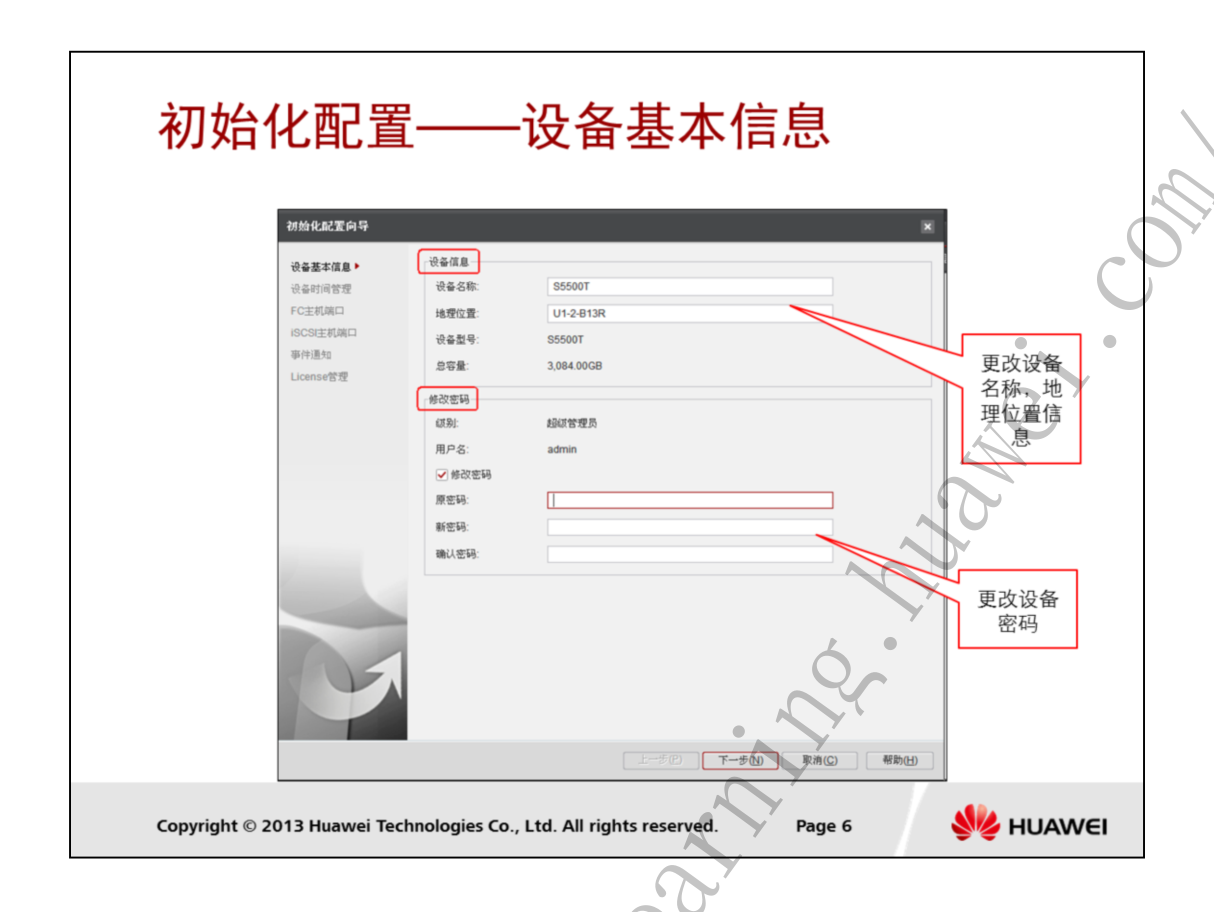
↓
结束

 Page 4 HUAWEI

- 阵列配置过程严格按照流程顺序进行
- 初次配置结束并且确认无误后，请使用信息收集工具或通过ISM保存配置信息
- 如果业务运行后需要修改配置，请使用信息收集工具或通过ISM保存配置信息



- 在ISM客户端发现设备后，可以通过初始化配置向导对存储设备进行相关设置，包括设备基本信息设置、设备时间设置、FC端口信息设置、iSCSI端口信息设置、事件通知设置、License管理 。



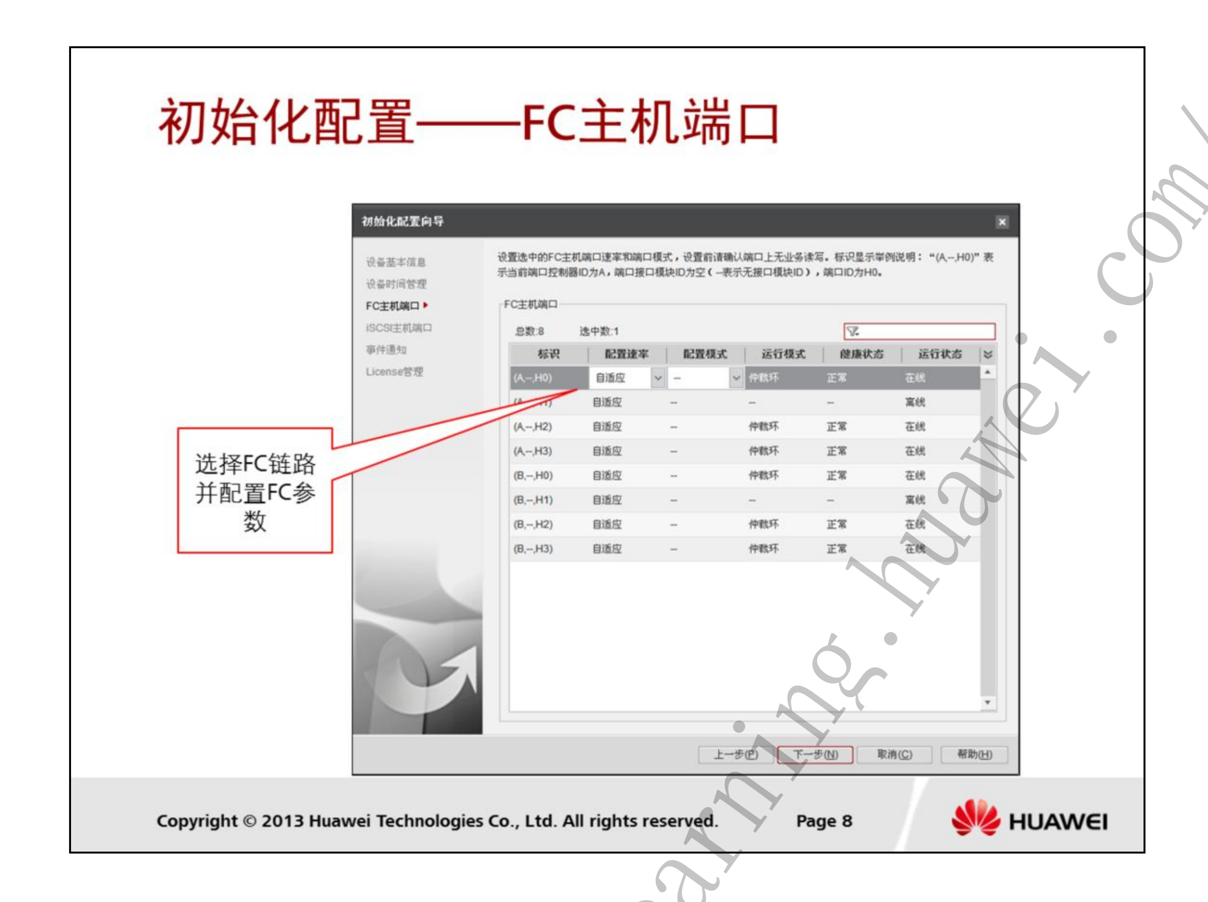
- 设备名称
- 只能包含半角的字母、半角的数字、“.”、“_”、“-”和简体中文字符，且“-”不能作为首字符，长度为1～32个字符（1个中文字符占3个字符长度）。
- 地理位置
- 只能包含半角的字母、半角的数字、空格、“.”、“_”、“-”和简体中文字符，且“-”和空格不能作为首字符，长度为1～32个字符（1个中文字符占3个字符长度）。
- 密码
  - 密码设置需要满足如下要求：密码为8到16位的字符串。
  - 密码必须包含如下至少两种字符的组合:
    - 一个或一个以上的小写字母。
    - 一个或一个以上的大写字母。
    - 一个或一个以上的数字。
    - 一个或一个以上的特殊字符：`，~，!，@，#，$，%，^，&，*，()，-，_，=，+，\，|，[，{，}，]，;，:，'，"，,，<，.，>，/，?和空格。
  - 密码不能和用户名或者用户名的倒写一样。

初始化配置——设备时间管理

修改设备时间

Page 7
HUAWEI

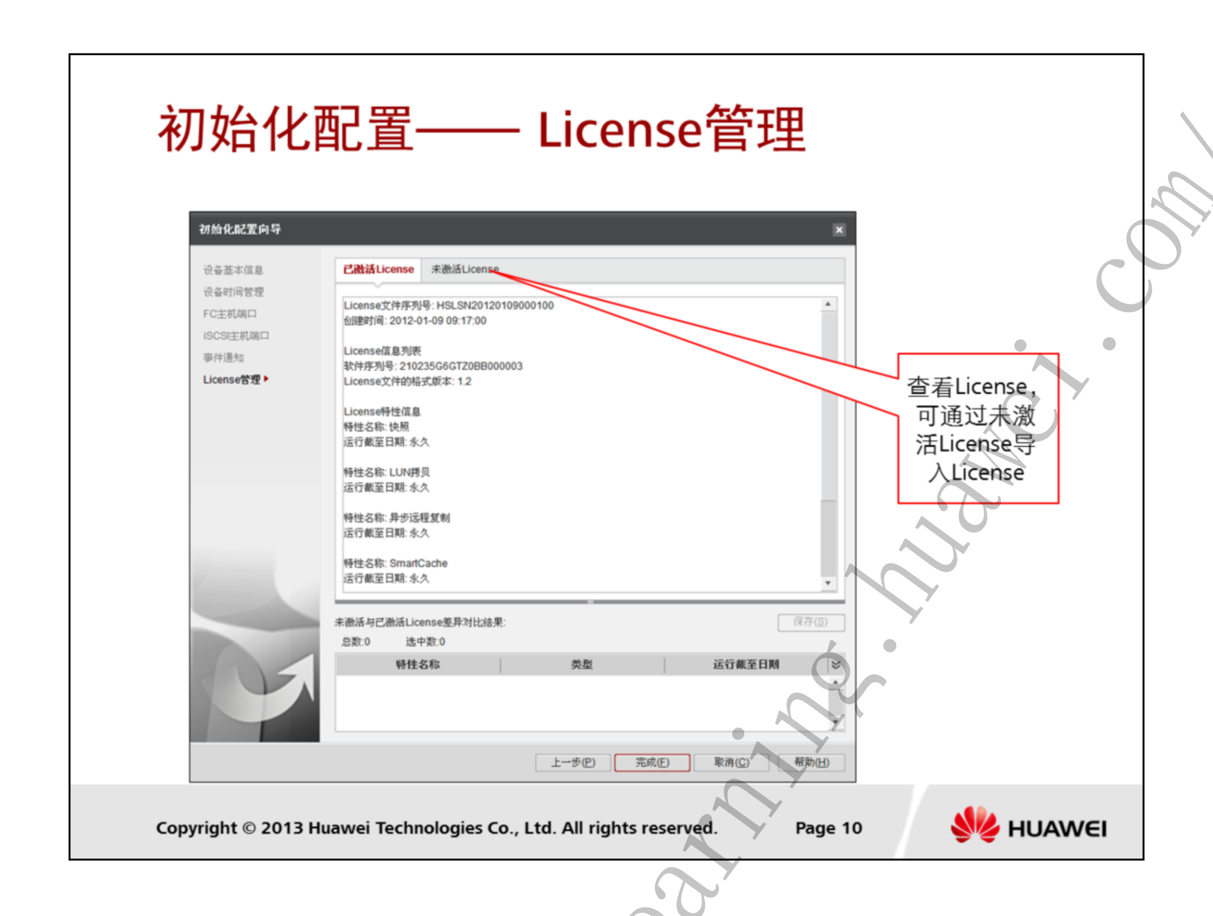
- 可以通过以下三种方式设置设备时间。
  - 选中“同步客户端时间”，同步当前客户端的时间和时区到设备。
  - 选中“NTP自动同步”，同步NTP（Network Time Protocol）服务器的时间到设备
    - 在“服务器IP地址”文本框输入NTP服务器的IP地址。 服务器IP地址支持IPv4地址和IPv6地址。
    - 在“同步周期（小时）”文本框中设置同步服务器时间到设备的周期。 取值范围为12～240的整数。
  - 选中“手动修改”，手动修改设备的时间和时区。
    - 在“设备时间”后单击“修改”。 系统弹出“修改时间”对话框。
    - 在“日期”区域框修改设备的显示日期。
    - 在“时间”区域框修改设备的显示时间。
    - 在“当前时区”下拉列表框选择设备所在的时区。
    - 单击“确定”。 系统弹出“提示”对话框。
    - 请仔细阅读对话框中的内容，确认后单击“确定”。系统弹出“信息”对话框，显示“操作成功”， 完成修改设备时间的操作



- FC主机端口速率
  - 该参数表示存储设备上FC主机端口的速率。
- 对系统的影响
  - 主要包括以下几个方面： FC主机端口速率会影响性能，设置的速率过低，导致性能达不到预期值。
  - 该参数设置错误，会导致应用服务器与存储设备之间通信失败。
  - 自适应、2Gbit/s、4Gbit/s和8Gbit/s。
- 取值原则
  - 该参数必须和应用服务器FC HBA卡端口的速率设置一致，否则会造成通信失败。
  - 在设置FC主机端口的速率时，要考虑FC HBA卡端口的速率。如果FC HBA卡的速率设置为“自适应”，则存储设备上FC主机端口的速率也可以设置为“自适应”；但不能设置为主机上FC HBA卡不支持的速率值。
- FC端口模式
  - 当存储设备通过光纤交换机与应用服务器连接时，建议选择交换机模式。
  - 当存储设备通过FC主机端口与主机或其他阵列直连时，建议选择点对点模式
  - 当存储设备与应用服务器的端口处于同一个仲裁环网时，建议选择仲裁环模式



- 选择iSCSI链路并配置IP地址以及路由信息,配置的IP地址在服务器端能够Ping通。



- 通过License管理功能，可以浏览License信息、导入并激活License文件。
- License是控制产品使用范围、功能、期限的许可文件。在以下情况下，存储系统需要导入相应的License文件：
  - 使用SmartCache功能；
  - 使用虚拟快照功能；
  - 使用LUN拷贝功能；
  - 使用同步远程复制功能；
  - 使用异步远程复制功能；
  - 使用分裂镜像功能；
  - 使用自动精简配置功能。

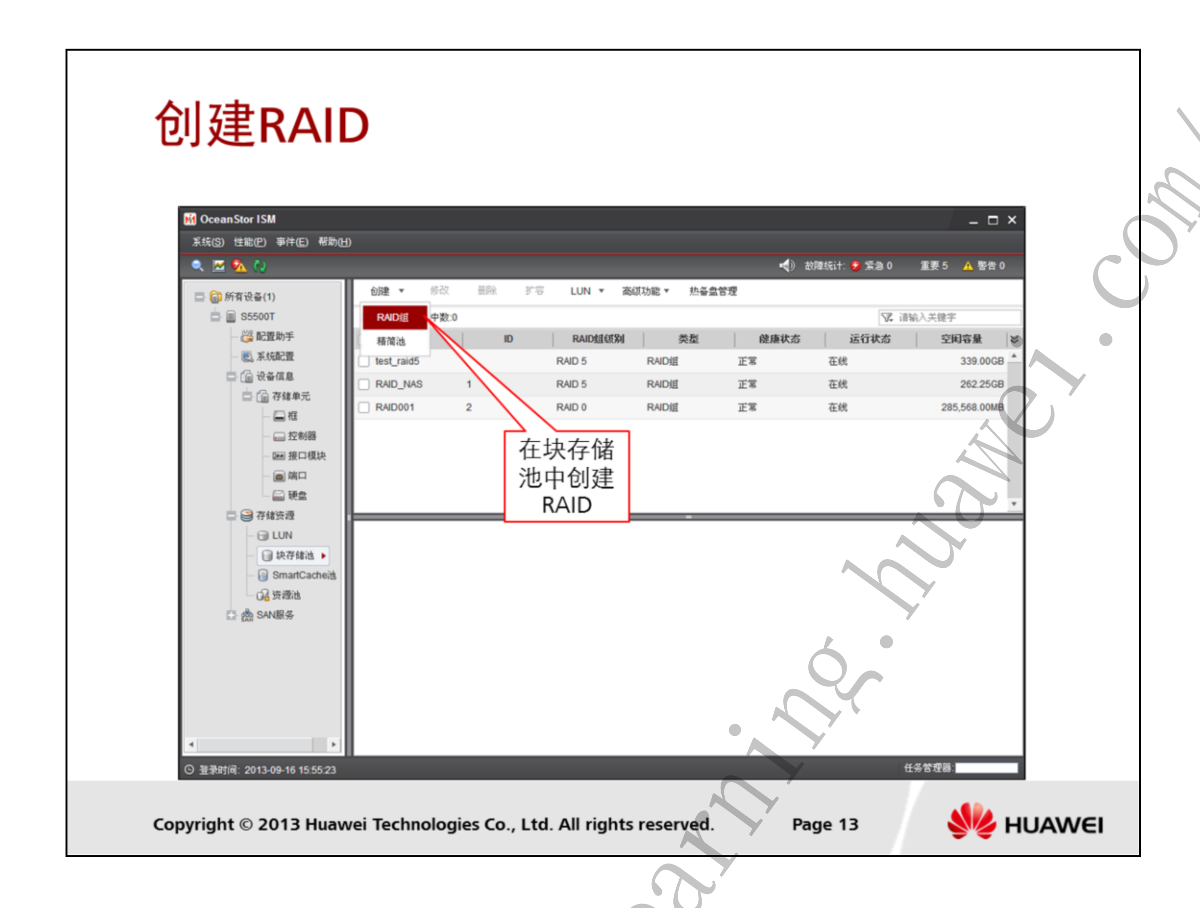


- 邮件通知
  - 通过该操作，系统以电子邮件的形式将选定级别的告警信息发送至指定的邮箱中。
- 短消息通知
  - 通过该操作，可以将指定级别的告警信息以短消息的形式发送至指定手机。
- 系统状态通知
  - 通过该操作，设置一个发送周期，邮件和短消息会按照设定的周期被发送到远程维护中心，便于监控当前的系统状态。
- Syslog通知
  - 通过该操作，可以将指定级别的系统日志保存到IP地址指定的服务器的特定文件夹中，供第三方软件调用，对存储系统的运行状态进行监控。



目录





- 存储设备支持RAID 0、RAID 1、RAID 3、RAID 5、RAID 6、RAID 10和RAID 50，可以根据不同的应用创建不同级别的RAID组，各种RAID组适用的应用场景如下所示：
  - RAID 0：适用于以读写性能为第一要求，且数据保护要求最小的应用。
  - RAID 1：适用于以数据安全为第一要求的应用。
  - RAID 3：适用于对存储性能、数据安全和存储成本兼顾的应用。RAID组中的一个硬盘作为校验盘。任意一个成员盘失效，RAID组降级仍能正常工作。
  - RAID 5：适用于对存储性能、数据安全和存储成本兼顾的应用。RAID组中存在分散在不同条带上的奇偶校验数据。任意一个成员盘失效，RAID组降级仍能正常工作。
  - RAID 6：适用于对数据安全要求较高的应用。任意两个成员盘失效，RAID组降级仍能正常工作。
  - RAID 10：适用于既有大量数据需要存取，同时又对数据安全性要求严格的应用，如银行、金融、商业超市、仓储库房、各种档案管理等。
  - RAID 50：适用于既有大量数据需要存取，同时对数据安全性要求较高的应用。RAID 50是RAID 5与RAID 0两种技术的结合，至少需要6个空闲盘。
- 创建RAID组时，RAID组成员盘的类型应保持一致。
- 如果存储系统中存在SATA硬盘和NL SAS硬盘时，可以选择SATA硬盘和NL SAS硬盘作为同一个RAID组的成员盘。



- 热备盘管理
  - 热备盘是被指定用于替代RAID组内故障成员盘的硬盘，完成的任务是承载故障硬盘中的数据。通过热备盘管理可以将空闲盘设置为热备盘或将热备盘设置为空闲盘。
- 设置热备盘
  - 在导航树上，展开存储设备下的“存储资源”节点。
  - 单击“块存储池”节点。
  - 在右侧信息展示区选择“热备盘管理”。系统弹出“热备盘管理”对话框。
  - 在“空闲盘”区域框中，勾选需要设置为热备盘的空闲盘。
  - 单击图标，空闲盘移动到“热备盘”区域框中。
  - 单击“确定”。系统弹出“执行结果”对话框，提示操作成功。
  - 单击“关闭”。
- 取消热备盘
  - 在导航树上，展开存储设备下的“存储资源”节点。
  - 单击“块存储池”节点。
  - 在右侧信息展示区选择“热备盘管理”。系统弹出“热备盘管理”对话框。
  - 在“热备盘”区域框中，勾选需要取消的热备盘。

创建LUN

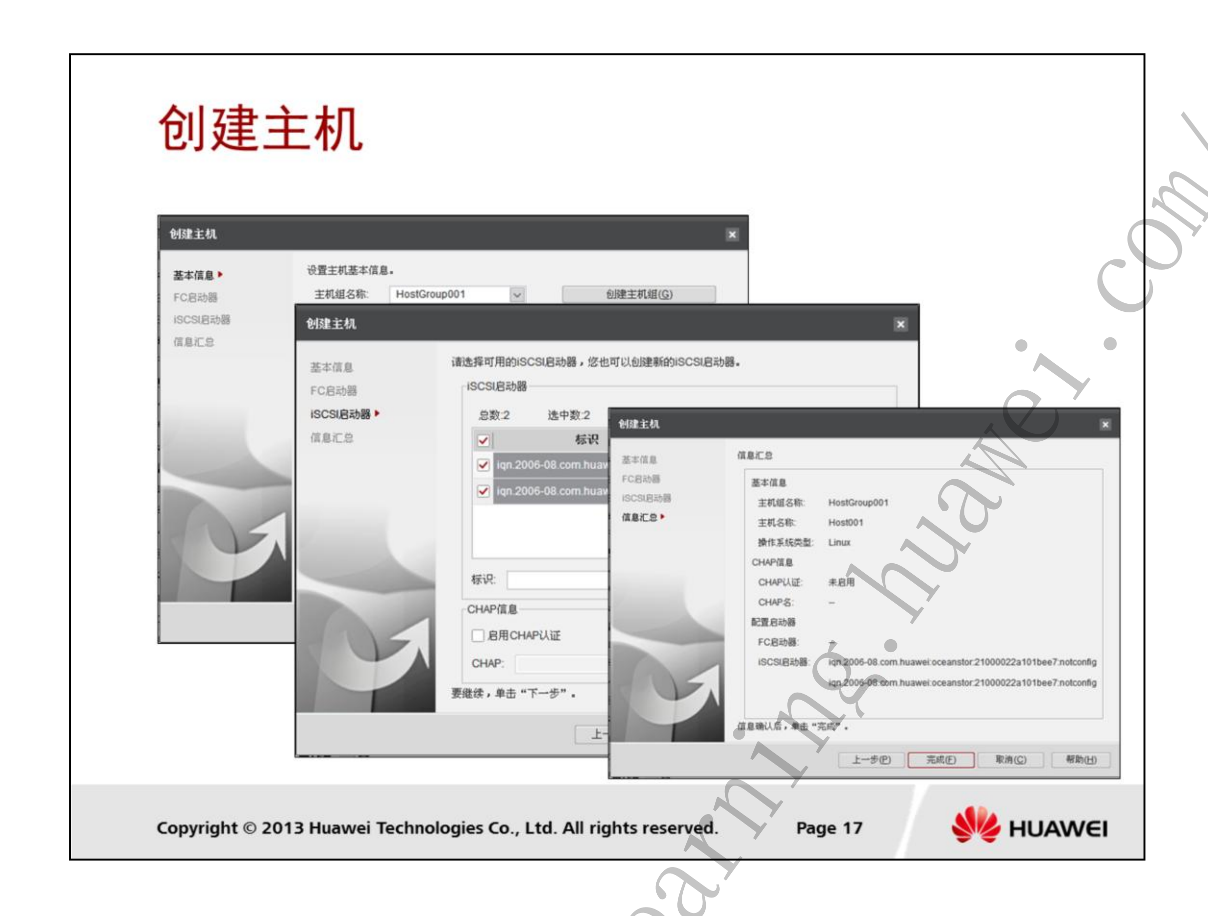
创建LUN的时候要考虑好分条深度的大小，创建好后不能再修改

 Page 15 HUAWEI

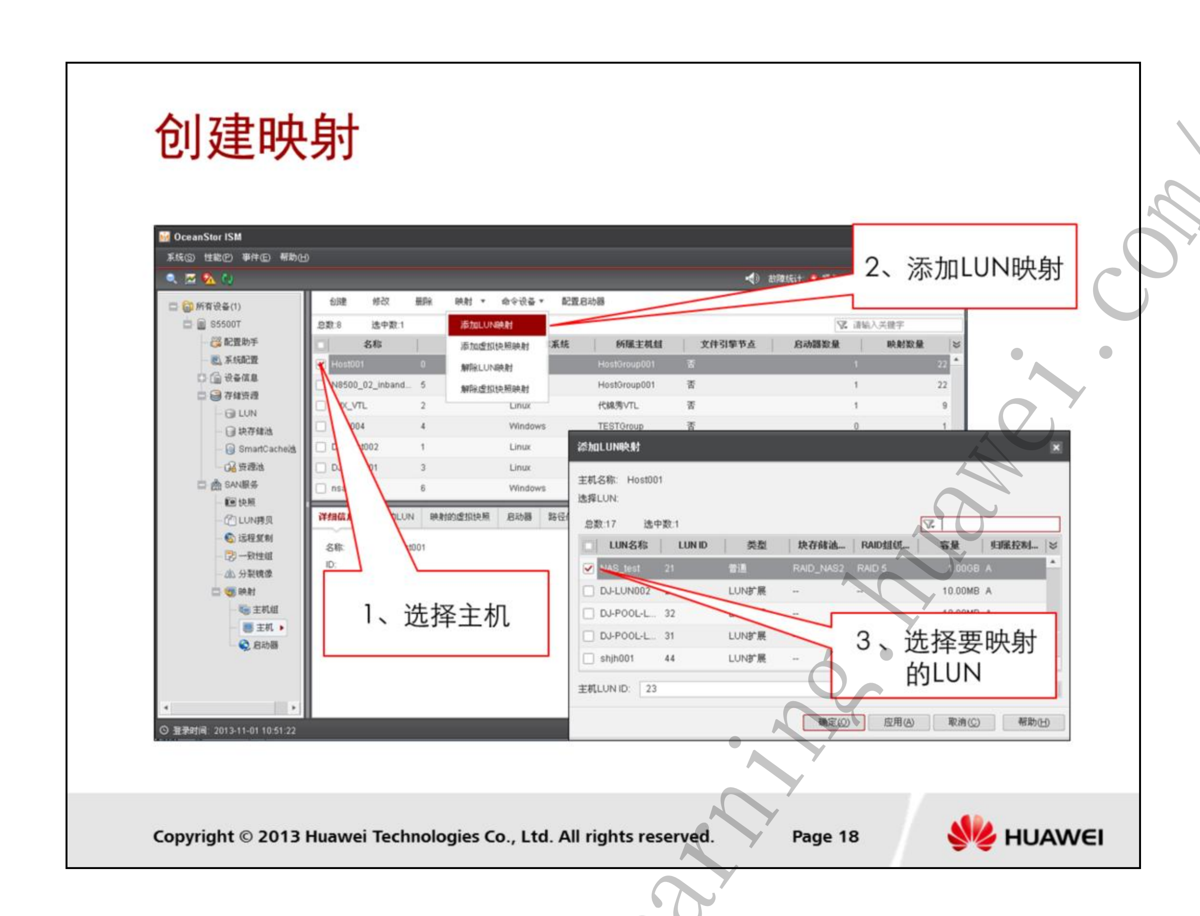
- 名称
  - 名称不能重复，只能包含半角的字母、半角的数字、“.”、“_”、“-”和简体中文字符，且“-”不能作为首字符，长度为1～32个字符（1个中文字符占3个字符长度）。
- 分条深度
  - 新建LUN的分条深度。分条深度是指在使用分条数据映射的硬盘设备中，条带内的块大小。也指在硬盘设备的单个成员盘区中，连续编址的虚拟硬盘块映射到连续编址的块的大小。分条深度取值包括“4 KB”、“8 KB”、“16 KB”、“32 KB”、“64 KB”、“128 KB”、“256 KB”和“512 KB”8种。
  - 由于分条深度影响到I/O性能，因此在不同的应用场景下应适当选择分条深度大小。例如：当系统应用于存储顺序数据较多的情况下，如存储媒体数据，建议设置较大的分条深度。推荐设置为64 KB。当系统应用于存储随机数据较多的情况下，如存储事务处理数据，建议设置较小的分条深度。推荐设置为32 KB。
- 归属控制器
  - 为了实现控制器A和控制器B的负载均衡，建议将存储系统中的LUN分配给不同的控制器。 说明： 当LUN所属RAID组的成员盘为SAS盘或FC盘时，LUN的归属控制器可以选择为“控制器A”、“控制器B”或“自动选择”。当LUN所属RAID组的成员盘为其他类型时，LUN的归属控制器只可以选择“控制器A”或“控制器B”。



- 在“主机组名称”文本框中输入新创建主机组的名称。
  - 名称不能与已有的主机组名称重复。
  - 只能包含半角的字母、半角的数字、“.”、“_”、“-”和简体中文字符，且“-”不能作为首字符。
  - 长度为1～32个字符（1个中文字符占3个字符长度）。



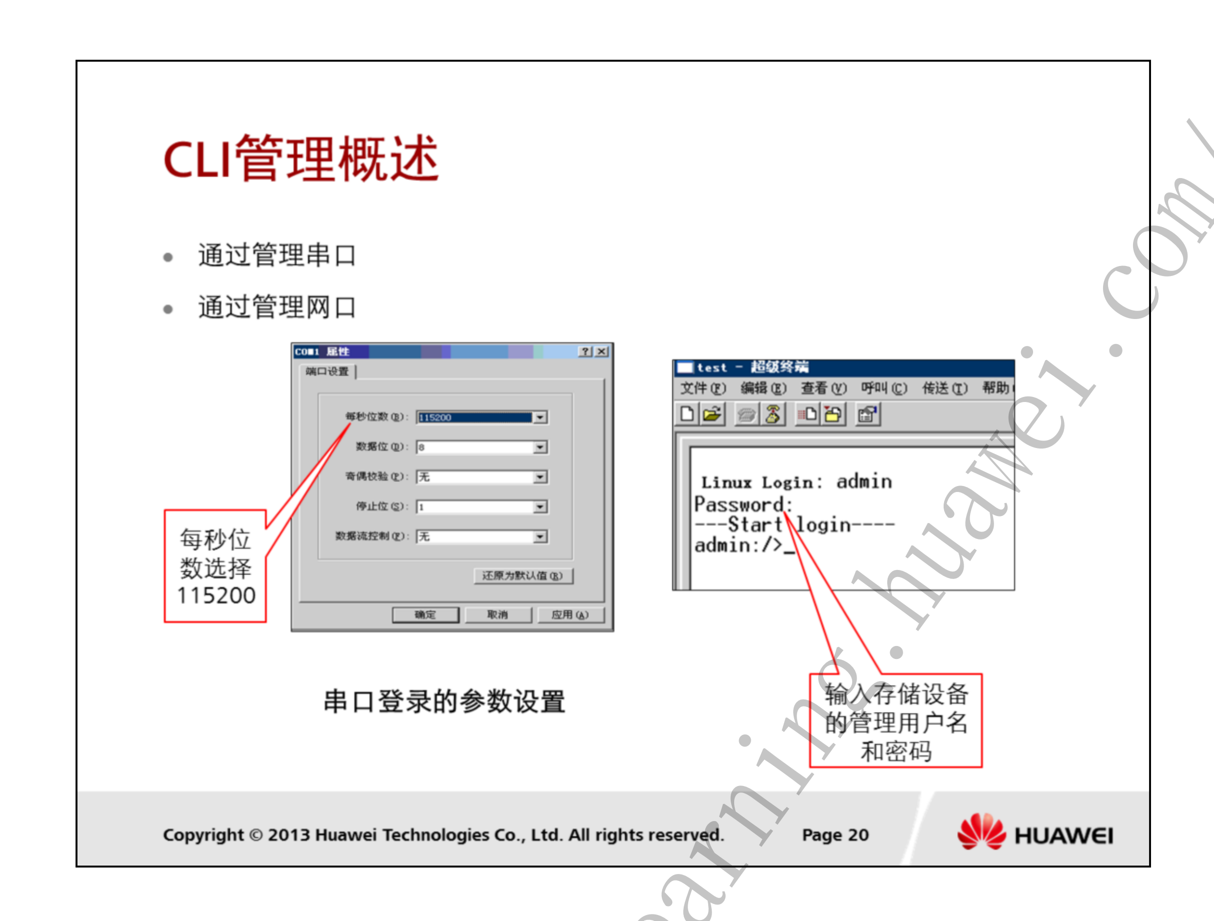
- 在“主机名称”文本框中输入新创建主机的名称。
  - 名称不能重复。
  - 只能包含半角的字母、半角的数字、“.”、“_”、“-”和简体中文字符，且“-”不能作为首字符。
  - 长度为1～32个字符（1个中文字符占3个字符长度）。
- 在“操作系统类型”后的下拉列表中选择主机的操作系统类型。



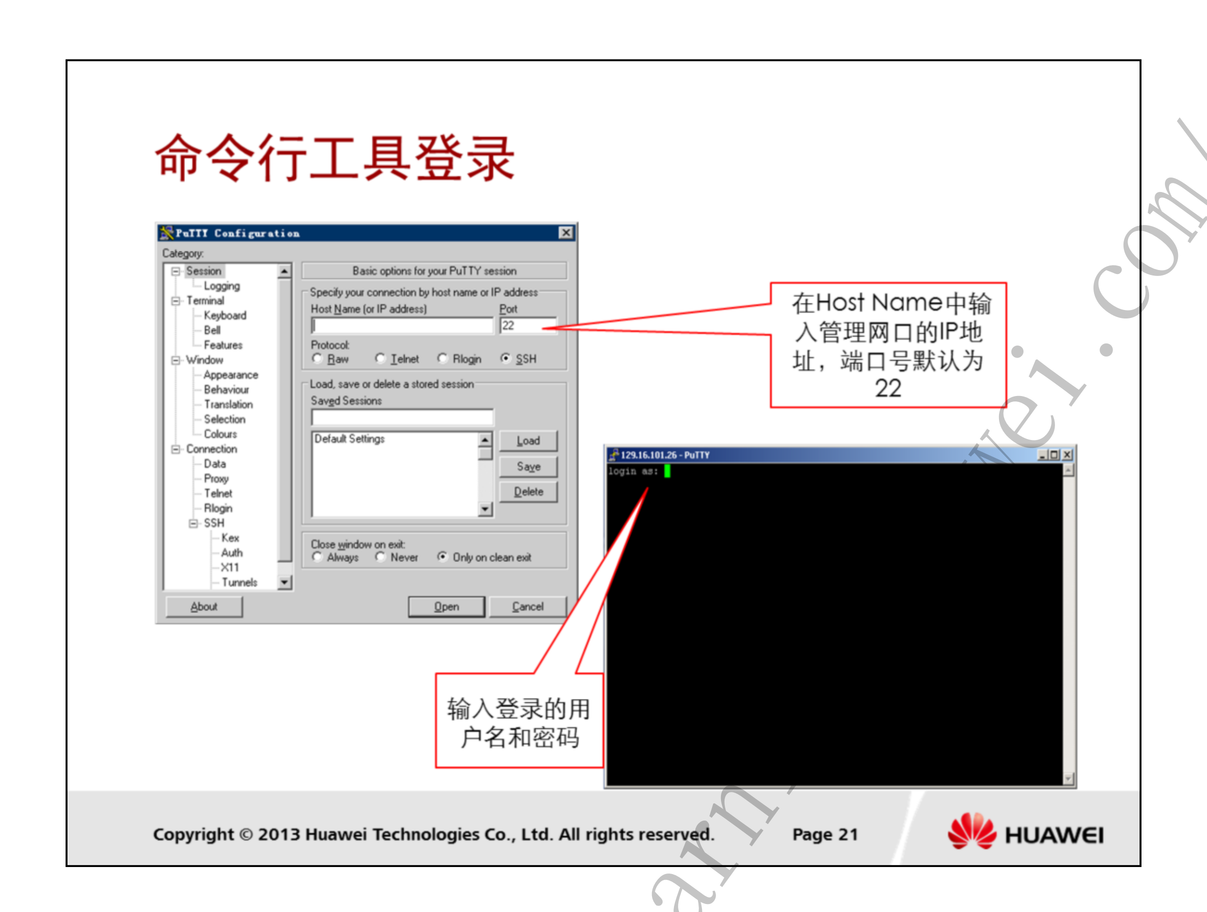
- 私有LUN不能映射给主机。
- LUN不能映射给默认主机组下的主机。
- 创建LUN映射可以从存储资源->LUN里面去操作配置，也可以通过映射主机组->主机->添加LUN映射操作。



- 配置启动器可通过“映射”->“主机组”->“主机”选项进去配置，也可以通过“映射”->“启动器”配置。



- 在不知道设备管理IP的时候，可以通过串口进入管理界面，查看管理IP。连接管理串口需要使用DB9转RJ45的串口线，设备的管理串口的具体位置参考产品手册。



- 常用的命令行工具有putty和ssh client。

常用命令

- showctrlinfo　查看控制器信息
- showsys　查看系统信息
- Showctrlip　查看管理IP
- showallver　查看版本信息
- chgctrlip　修改管理IP
- exit　系统退出命令

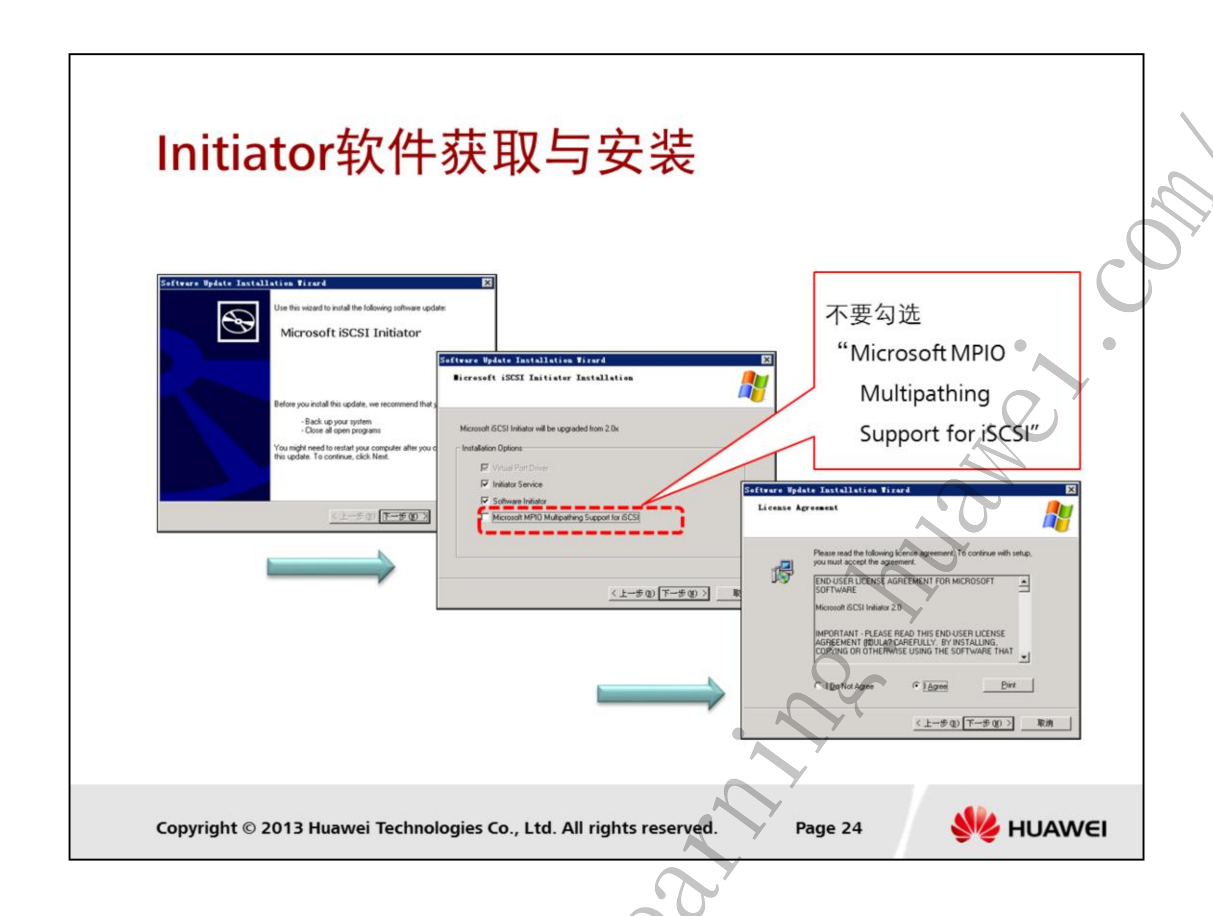


- 可以使用tab键补全命令
- 可以在命令行后添加help查看帮助信息



目录





- 在Windows应用服务器上通过“Microsoft iSCSI Initiator”软件配置目标器的IP地址、用户名和密码，可以建立存储设备和应用服务器的连接。
  - “Microsoft iSCSI Initiator”安装程序可从微软网站
  - http://www.microsoft.com上下载。请使用2.01及以上版本。下载完成后双击安装程序即可开始安装。
  - 如果应用服务器和存储的连接存在多条路径，需要安装华赛研发的UltraPath软件，在安装“Microsoft iSCSI Initiator”软件时，不要勾选“Microsoft MPIO Multipathing Support for iSCSI”。
  - 安装完成后，请重新启动应用服务器。

Windows 平台下iSCSI服务操作

计算机管理

文件(F) 操作(A) 查看(V) 窗口(W) 帮助(H)

计算机管理(本地)
- 系统工具
  - 事件查看器
  - 共享文件夹
  - 本地用户和组
  - 性能日志和警报
  - 设备管理器
- 存储
  - 可移动存储
  - 磁盘碎片整理程序
  - 磁盘管理
- 服务和应用程序
  - 电话服务
  - 服务
  - WMI 控制
  - 索引服务

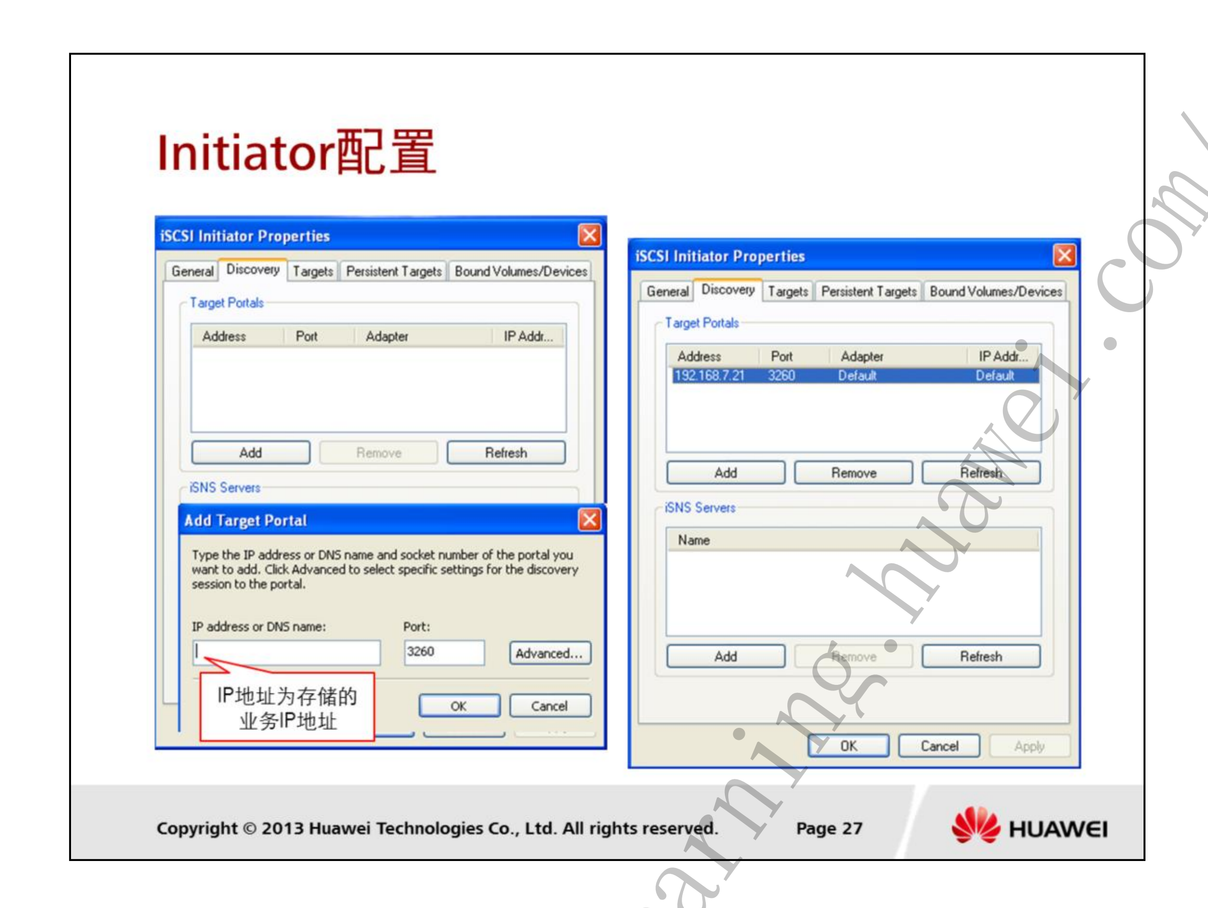
名称	描述	状态	启动类型	登录为
Kerberos Key Distribution Center	在...		禁用	本地系统
License Logging	监...		禁用	网络服务
Logical Disk Manager	监...	已启动	自动	本地系统
Logical Disk Manager Administrat...	配...		手动	本地系统
Messenger	传...		禁用	本地系统
Microsoft iSCSI Initiator Service	Man...	已启动	自动	本地系统
Microsoft Software Shadow Copy P...	管...		手动	本地系统
NemService			手动	本地系统
Net Logon	为...		手动	本地系统
Net.Tcp Port Sharing Service	Pro...		禁用	本地服务
NetMeeting Remote Desktop Sharing	使...		禁用	本地系统
Network Connections	管...	已启动	手动	本地系统
Network DDE	为...		禁用	本地系统
Network DDE DSDM	管...		禁用	本地系统
Network Location Awareness (NLA)	收...	已启动	手动	本地系统

 Page 25 HUAWEI

- 管理iSCSI Initiator服务
  - 在桌面左下角点击“开始”；
  - 在弹出的导航栏点击“管理工具”；
  - 点击“管理工具”下的“服务”，如图所示。
- iSCSI Initiator的服务名称为Microsoft iSCSI Initiator Service，如图红色虚框所表示。
- 双击Microsoft iSCSI Initiator Service，弹出的窗口如图所示。在此可以对该服务进行启动、停止等操作和配置。



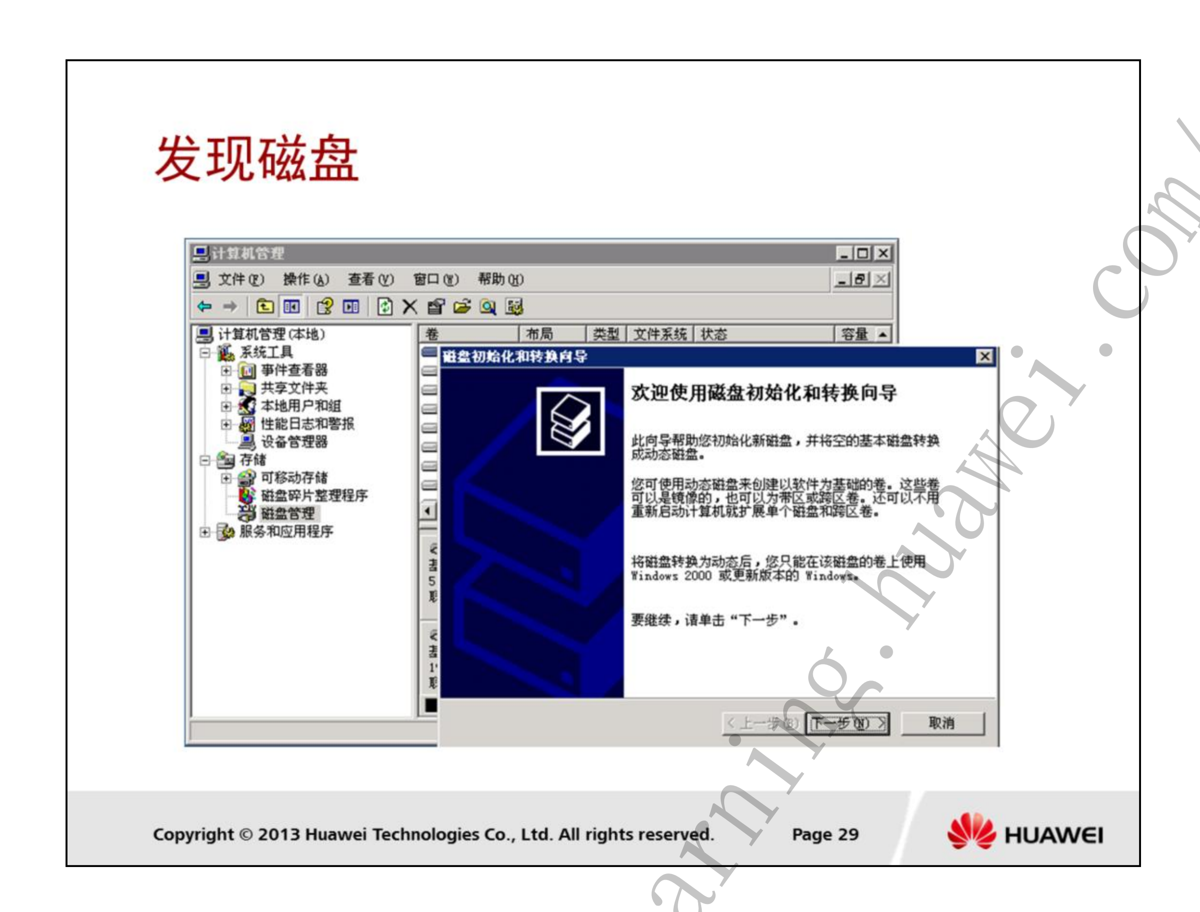
- 这里的启动器名称要配置在存储设备上的启动器配置，然后再在创建主机的时候，将该启动器配置给主机。建立存储上创建的主机到服务器之间的映射关系。



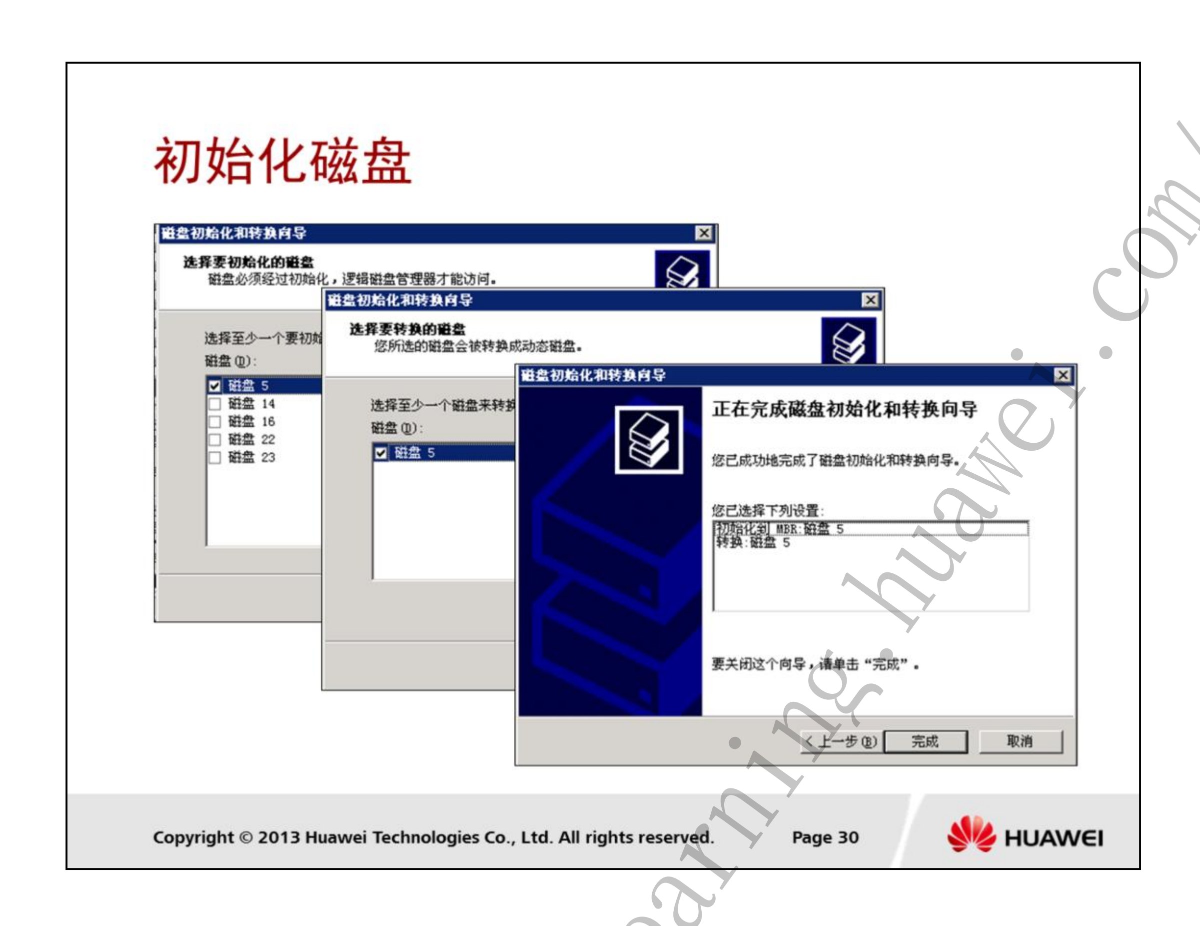
- 如果存储和服务器之间是通过FC HBA卡连接，不需要此配置步骤，FC能够自动发现并建立连接；
- 如果是通过IP网络连接，则要进行如图的IP地址配置，IP地址为存储设备的业务口IP地址。



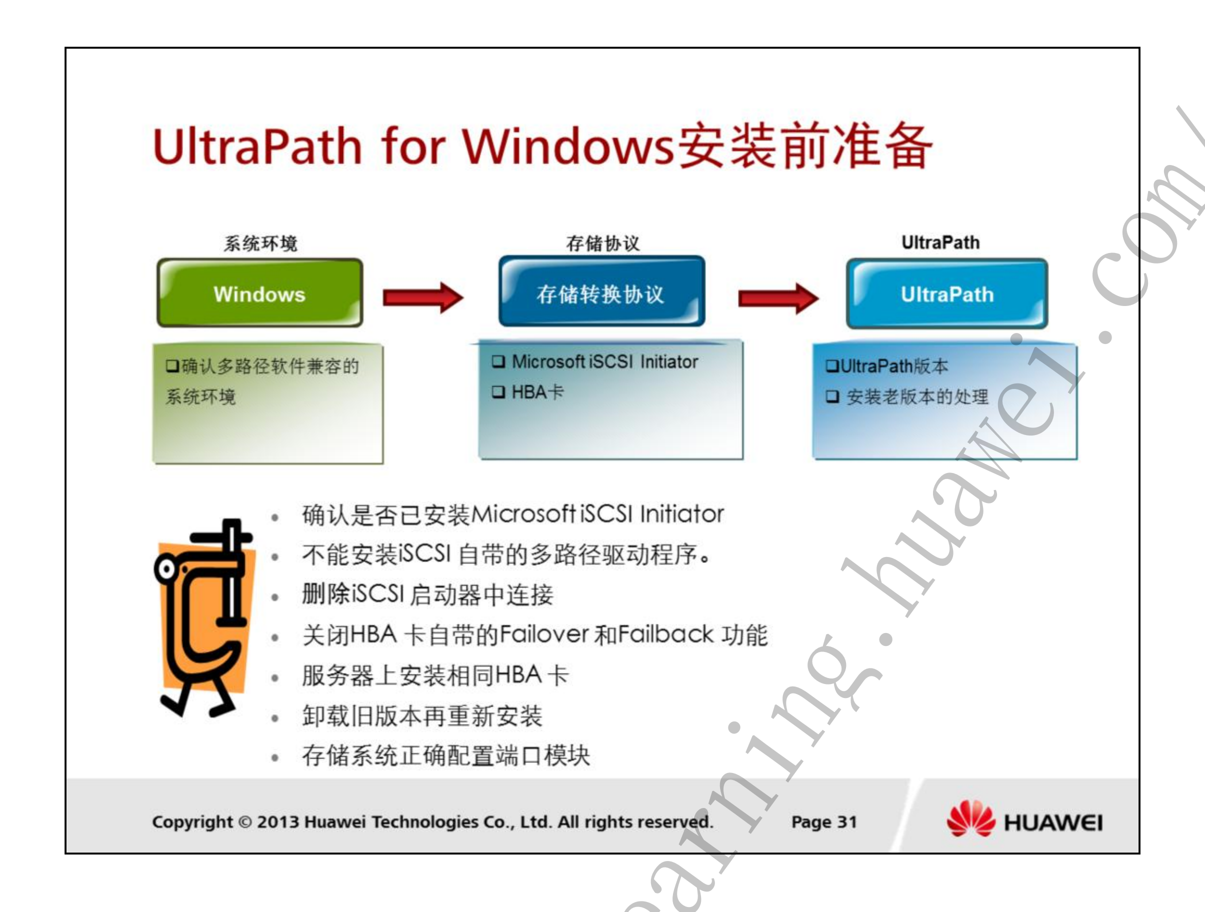
- 在建立服务器到存储之间iSCSI连接之前，需要在存储端配置好主机，映射LUN,并给主机指定服务器的启动器名称。击LogON,建立服务器到存储之间的连接。然后服务器就能发现磁盘。



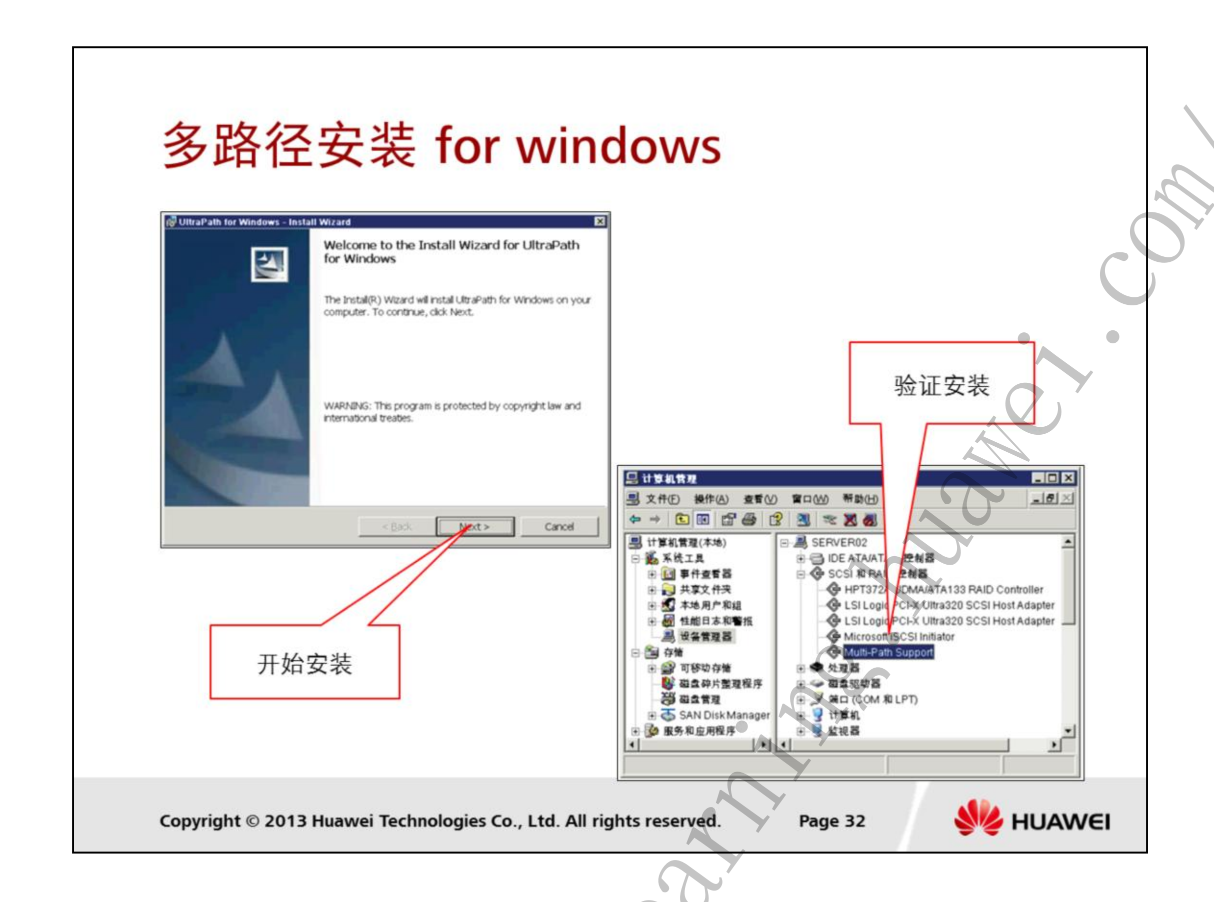
- 当建立连接后，打开计算机管理，进入磁盘管理，会自动跳出一个磁盘初始化向导。



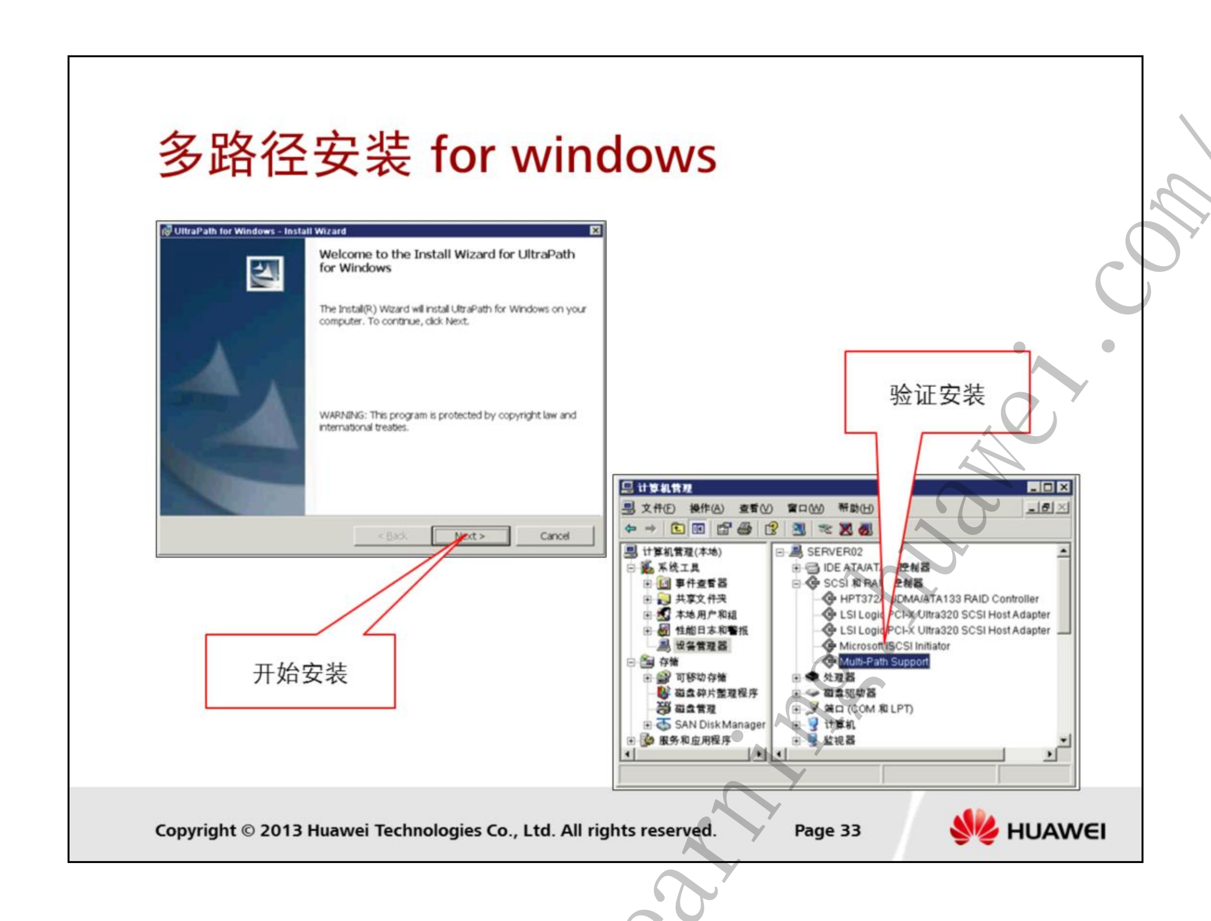
- 按照初始化向导，一步一步完成磁盘初始化操作，完成后就能在磁盘管理中发现磁盘。



- 步骤1 查询当前应用服务器的Windows 操作系统的版本及补丁是否满足安装UltraPath for Windows 的要求。
- 步骤2 确认使用于当前Windows 版本的UltraPath for Windows 安装包与应用服务器的Windows操作系统类型相匹配。
- 步骤3 确认应用服务器上没有安装老版本的UltraPath for Windows，如果已安装有老版本的UltraPath for Windows，请先正确卸载老版本的UltraPath for Windows。
- 步骤4 如果安装UltraPath for Windows 的应用服务器是Windows 2000 操作系统，请确认应用服务器上已安装Windows 2000 操作系统安装光盘附带的Support Tools。
- 步骤5 如果应用服务器上安装有Microsoft iSCSI Initiator，请参见相关材料确认应用服务器确认Microsoft iSCSI Initiator 是否满足安装UltraPath for Windows 要求。
- 步骤6 如果应用服务器上安装有FC HBA 卡，请参见相关文档确认HBA卡是否满足安装UltraPath for Windows 要求。
- 多路径软件获取方式：
  - 1、华为官方网站下载；
  - 2、随机光盘；
- 多路径软件兼容性查询方式：
- 在发布的多路径下载软件包中，有对该版本多路径软件兼容性的说明，请仔细检查。



- 步骤1　将UltraPath for Windows 的安装程序拷贝到应用服务器上
  - 在安装UltraPath for Windows 前，请先断开对映射磁盘的一切应用。
  - 在安装UltraPath for Windows 前，先断开应用服务器与存储系统的连接。
- 步骤2 根据操作系统类型选择安装程序，运行可执行文件，系统进入UltraPath for Windows 安装界面。如左上图所示。
- 步骤3 单击“Next”，系统进入“UltraPath for Windows – Install Wizard”界面。
- 步骤4 在“User Name”、“Organization”中分别输入用户和组织名称。
- 步骤5 单击“Next”，系统进入“Ready to Install the Program”界面。
- 步骤6 单击“Install”，系统进入“Installing UltraPath for Windows”界面，显示安装进度。如果没有设置忽略驱动程序签名，系统会弹出“安全警报 – 驱动程序安装”对话框。单击“是”。
- 步骤7 安装完成，如左下图所示。
- 步骤8 单击“Finish”，系统弹出“UltraPath for Windows Installer Information”对话框。
- 步骤9 单击“Yes”，应用服务器重启。
- 步骤10 查看UltraPath for Windows 是否安装成功。
  - 单击“开始”，选择“所有程序 > UltraPath for Windows > Launch。



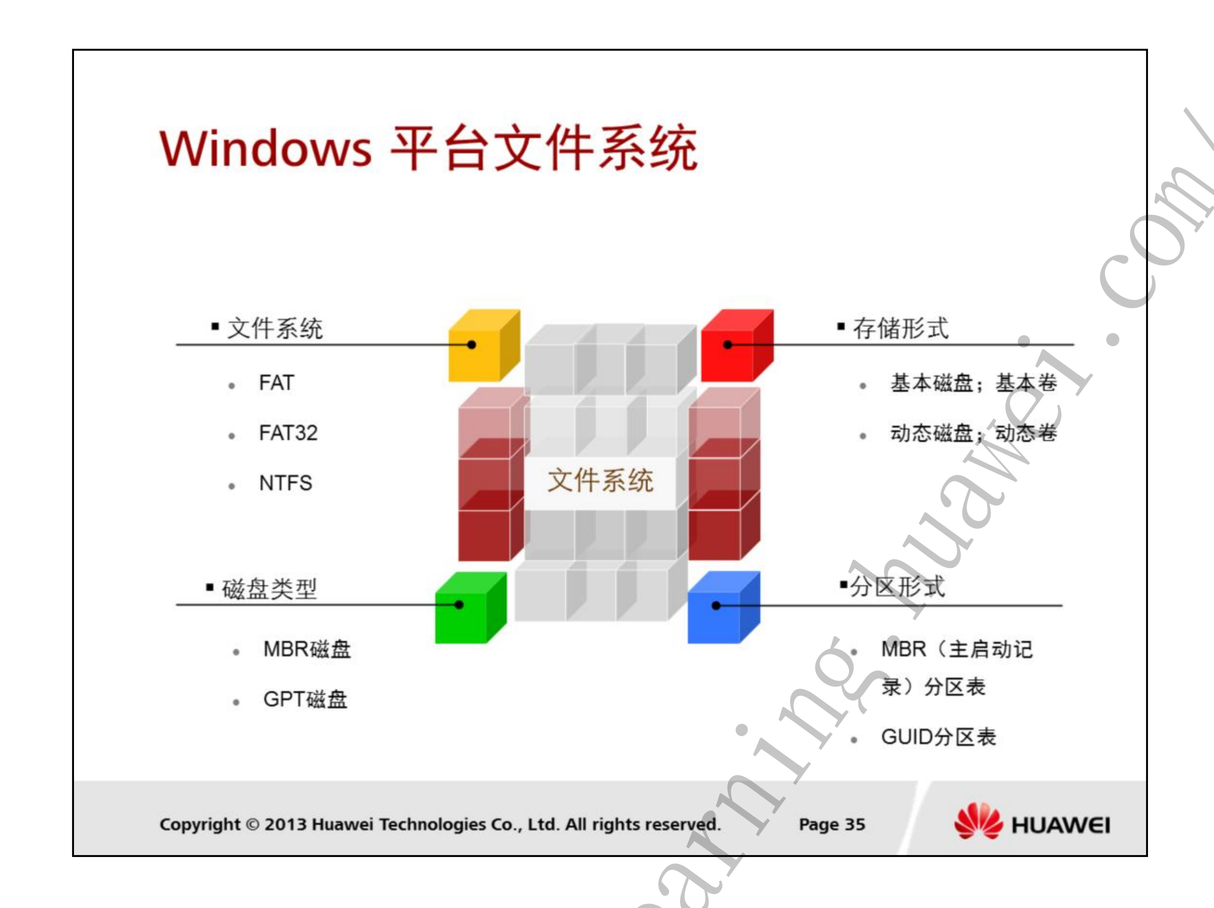
- UPManager.exe”，查看UltraPath for Windows Manager 工具。
- 也可以执行以下步骤查看UltraPath for Windows 是否安装成功。
    - 选中“我的电脑”图标后，单击鼠标右键，在弹出的快捷菜单中选择“管理”。
    - 在“计算机管理”界面左侧导航树中，选择“系统工具 > 设备管理器”。
    - 在界面右侧展开“SCSI 和RAID 控制器”节点，如果安装成功则显示“Multi-PathSupport”，如右图所示。

多路径安装 for windows

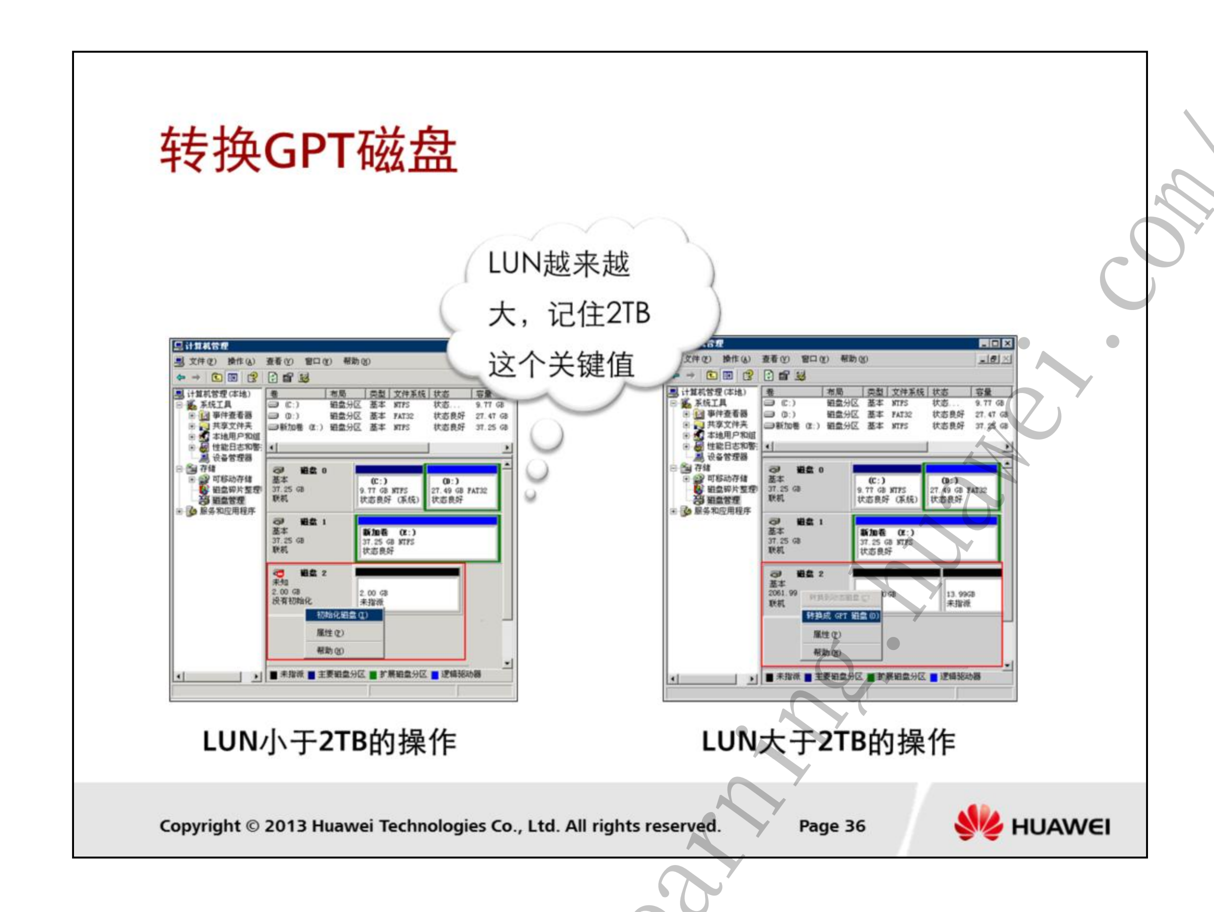
未安装多路径UltraPath

已安装多路径UltraPath

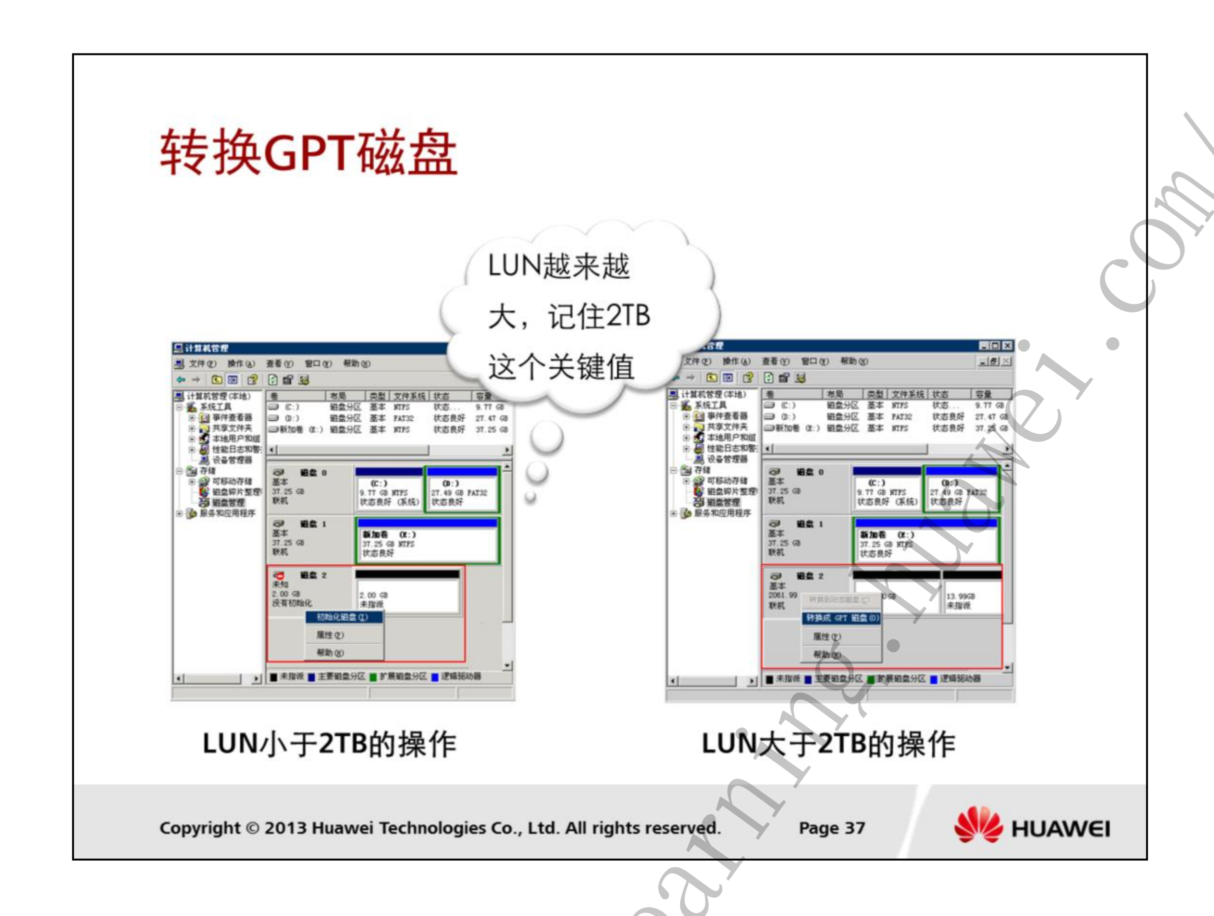
- 安装多路径软件UltraPath前磁盘状态:
  - 选中“我的电脑”图标后，单击鼠标右键，在弹出的快捷菜单中选择“管理”。
  - 在“计算机管理”界面左侧导航树中，选择“系统工具 > 设备管理器”。
  - 在界面右侧展开展开磁盘控制器。未安装多路径软件UltraPath时磁盘如左图所示，每一条应用服务器至LUN的链路均显示为一个磁盘；点击开始>管理工具>计算机管理>磁盘管理，可以看到磁盘数量亦与设备管理器所显示一致。
- 安装多路径软件UltraPath后，还需在应用服务器端扫描存储系统映射给应用服务器的LUN:
- 重复步骤1-3，选中“磁盘驱动器”，单击鼠标右键，在弹出的快捷菜单中选择“扫描检测硬件改动 ”。
- 系统扫描到由于冗余连接而产生的磁盘驱动器以及通过UltraPath for Windows 屏蔽掉LUN重影现象后的虚拟的LUN映射磁盘驱动器，如右图所示。



- 分区形式是指操作系统用于组织磁盘上的分区的方法。主要有MBR分区表和GUID分区表。
  - 主启动记录 (MBR) 磁盘分区：支持最大卷为 2 TB (terabytes) 并且每个磁盘最多有 4 个主分区（或 3 个主分区，1 个扩展分区和无限制的逻辑驱动器）
  - GUID 分区表 (GPT) 磁盘分区：支持最大卷为 18 EB (exabytes) 并且每磁盘最多有 128 个分区。
- 所有基于 x86 和基于 x64 的计算机都使用称为主启动记录 (MBR) 的分区形式。MBR 是未安装带有 Service Pack 1 (SP1) 的 Windows Server 2003 的基于 x86 的计算机上唯一可用的分区形式；基于 Itanium 的计算机、运行带有 Service Pack 1 (SP 1) 的 Windows Server 2003 的基于 x86 的计算机，以及基于 x64 的计算机，都可以使用 MBR 分区形式和 GUID 分区表 (GPT) 分区形式。“磁盘管理”将使用主启动记录分区形式的磁盘标记为 MBR 磁盘，而将使用 GUID 分区表分区形式的磁盘标记为 GPT 磁盘。
- 基本磁盘和基本卷，基本磁盘是包含主分区、扩展分区或逻辑驱动器的物理磁盘。基本磁盘上的分区和逻辑驱动器称为基本卷。只能在基本磁盘上创建基本卷。
  - 对于主启动记录 (MBR) 磁盘，可以最多创建四个主分区，或最多三个主分区加上一个扩展分区。在扩展分区内，可以创建多个逻辑驱动器。
  - 对于 GUID 分区表 (GPT) 磁盘，最多可创建 128 个主分区。由于 GPT 磁盘并不限制四个分区，因而不必创建扩展分区或逻辑驱动器。



- MBR分区方式仅仅支持每磁盘4个分区和最大2TB的磁盘容量。所以当映射给主机的LUN容量大于2TB时，需要将其转换成GPT磁盘才能正常使用。
    - 1、进入“计算机管理”对话框，在桌面上右键单击“我的电脑”图标，在弹出的快捷菜单中选择“管理”。
    - 2、在导航树上的“磁盘管理”节点下，扫描应用服务器上的新增的逻辑磁盘。
        - a)在“计算机管理”对话框的导航树上，选择“存储 > 磁盘管理”。
        - b)右键单击“磁盘管理”，在弹出的快捷菜单中选择“重新扫描磁盘”。
        - 扫描完成后在右侧的区域可以看到新增的逻辑磁盘（以“磁盘2”为例进行说明），如左图中红色区域所示（硬盘大小不同时显示稍有差异）。
    - 3、对新增的逻辑磁盘进行初始化。
        - a)选中“磁盘2”（如左图中红色区域所示），单击鼠标右键，并在弹出的快捷菜单中选择“初始化磁盘”。
        - b)在弹出的“初始化磁盘”对话框中，勾选需要初始化的逻辑磁盘并单击“确定”。
        - 等待1分钟，当“磁盘2”的状态变为“联机”时，初始化成功。
    - 4、当新增的逻辑磁盘大于2 TB时，请将逻辑磁盘转化为GPT磁盘，否则该逻辑磁盘将无法被访问。
        - a)选中“磁盘2”（以磁盘2为例），单击鼠标右键，并在弹出的快捷菜单中选择“转



  - 化成GPT磁盘”（如右图中红色区域所示）。
  - 转化成功后，逻辑磁盘将由两个分区转化为一个分区。
- 5、对逻辑磁盘进行分区并格式化。
  - 如果在Windows Server 2003图形界面下灰化了转化GPT磁盘的功能，可使用系统自带的diskpart.exe工具，转化之后可以对磁盘进行分区操作。

- 注：在磁盘分区不大于2TB的情况下，推荐使用基本磁盘形式创建磁盘分区。



目录



Initiator软件获取

- 对于iSCSI连接，安装多路径软件前需要先安装Initiator软件
- Open-iSCSI是用于RedHat Linux 5和SuSE Linux 10及更高版本的Linux系统与IP SAN设备对接的Initiator软件。
- 获取途径
  - 操作系统安装光盘
  - 官方网站：http://www.open-iscsi.org



- 获取Open-iSCSI软件的方法有：
  - 从安装光盘中查找open-iscsi软件；
  - 前往 http://www.open-iscsi.org/ 官方网站下载。
- Open-iSCSI在Linux下的配置，以SUSE 10.0为例：
  - 以root用户登录SUSE应用服务器。
  - 启动iSCSI服务。
  - #/etc/init.d/open-iscsi start
  - （可选）修改启动器的名称。

Initiator软件安装

Terminal
File Edit View Terminal Tabs Help
YaST @ linux-iu Press F1 for Help
YaST Control Center
Software
Hardware
System
Network Devices
Network Services
Novell AppArmor
Security and Users
Virtualization
Miscellaneous
NTP Configuration
Network Services (xinetd)
Proxy
Remote Administration
Routing
SLP Server
Samba Server
TFTP Server
WOL
Windows Domain Membership
iSCSI Initiator
iSCSI Target
[Help] [Quit]

 Page 40 HUAWEI

- Open-iSCSI在Linux下的安装，以SUSE 10.0为例：
  - 输入命令yast，弹出“YaST”界面。
  - 在界面左侧选择“Software”，移动光标在界面右侧选择“Software Management”，按“Enter”。
  - 按快捷键“Alt+S”搜索所有包含文字“iSCSI”的文件，按快捷键“Alt+O”确定操作。
  - 选择安装open-iscsi，按快捷键“Alt+A”确定操作。
  - 按照系统提示操作正确安装rpm 包。

Initiator软件配置登录

- 配置目标器的IP地址并登录目标器。
  - 运行iscsiadm -m discovery -t st -p XXX.XXX.XXX.XXX命令添加目标器。 以iSCSI主机端口的IP地址为192.168.123.100为例进行描述。

    # iscsiadm -m discovery -t st -p 192.168.123.100
- 运行iscsiadm –m node –l命令登录目标器。

    # iscsiadm –m node –l

 Page 41 

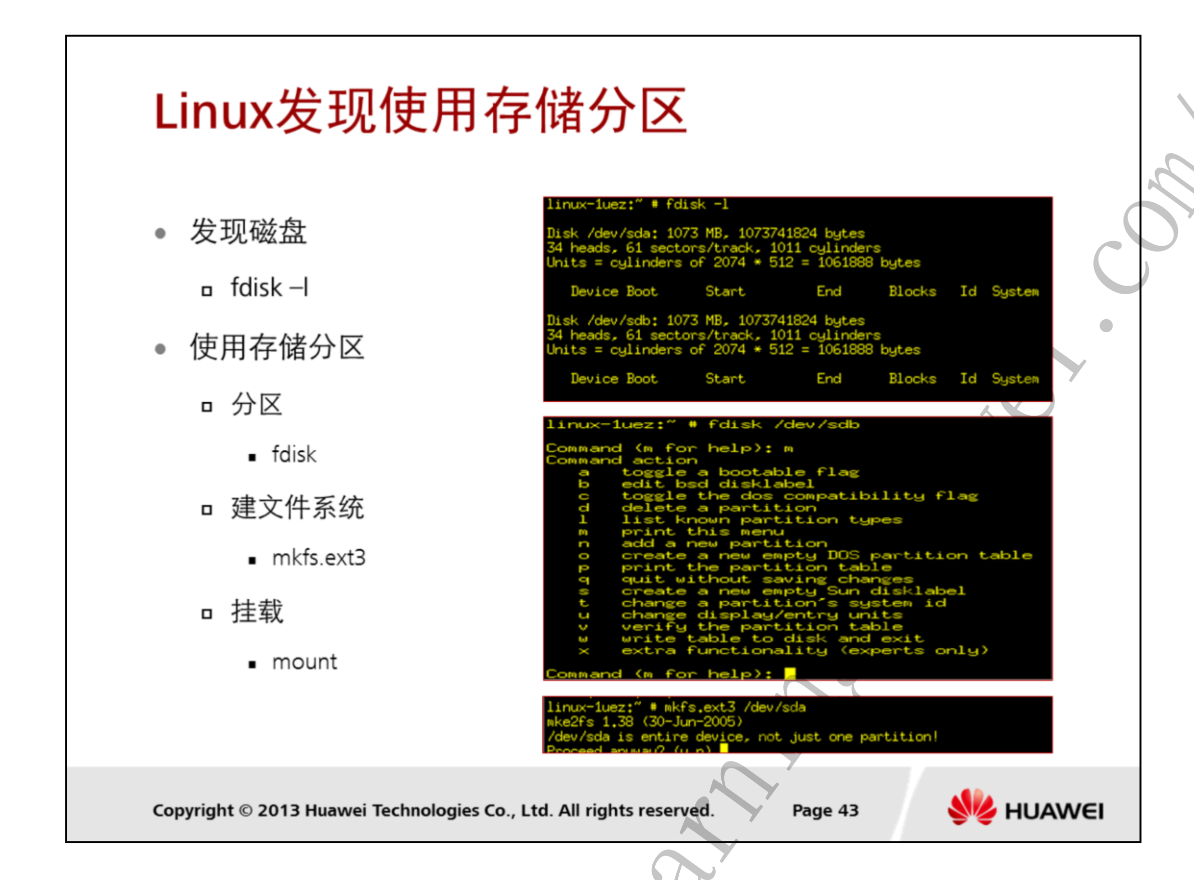
- 设置应用服务器通过此目标器端口持续访问存储设备。
- 在/etc目录下，运行vi iscsid.conf命令，打开“iscsid.conf”文件。
- 按i键进入编辑模式，编辑“iscsid.conf”文件。
- 将node.start设置为“automatic”。输入如下信息：
- node.startup=automatic
- 如果将node.startup设置为“manual”，重新启动应用服务器后，需重新建立基于此目标器端口的iSCSI连接。
- 按“ESC”键退出编辑模式。
- 输入:wq命令并按“Enter”键，保存并退出“iscsid.conf”文件。
- 重启iSCSI服务使配置生效。 运行open-iscsi restart命令重启open-iscsi服务。
- # /etc/init.d/open-iscsi restart

Linux平台open-iscsi服务操作

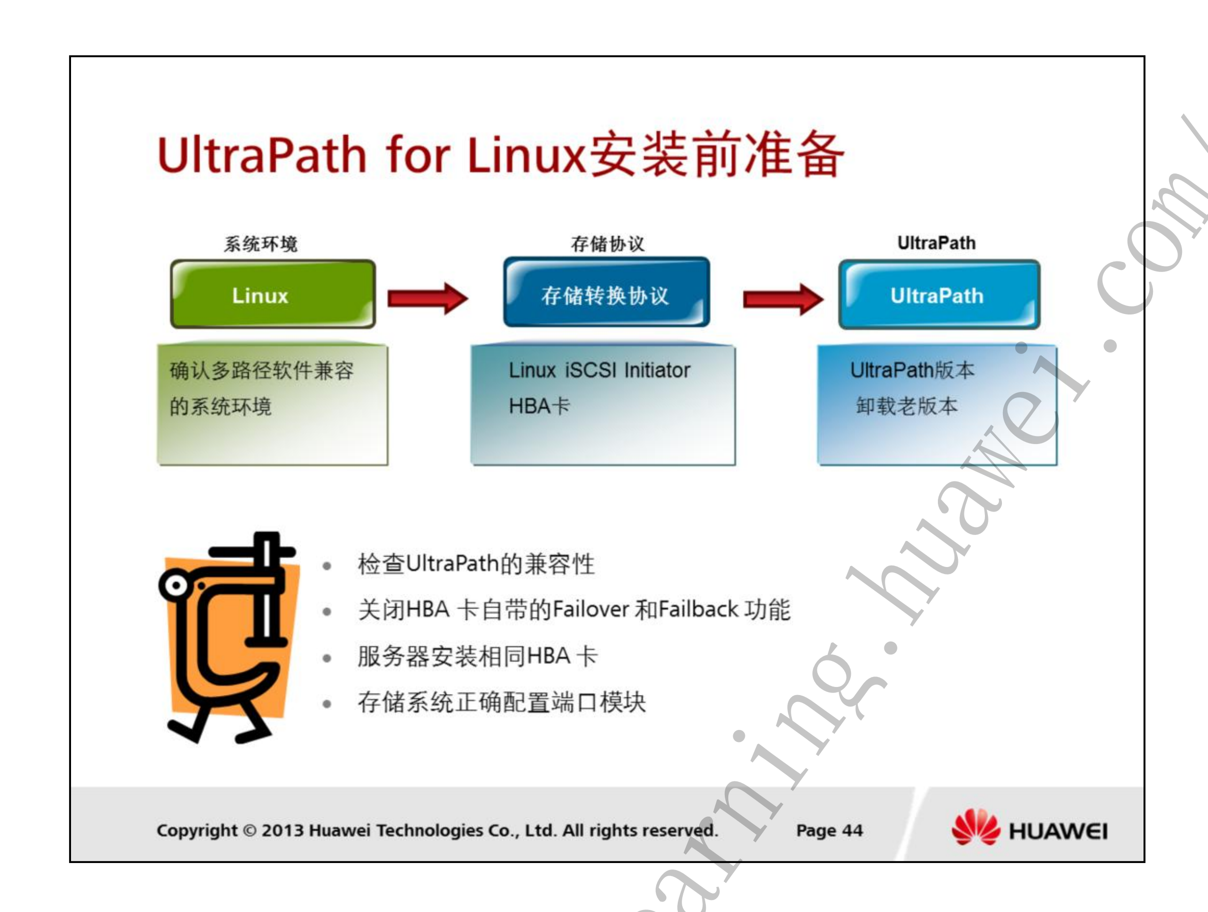
- 停止open-iscsi服务
  - # service open-iscsi stop
- 启动open-iscsi服务
  - # service open-iscsi start
- 重启open-iscsi服务
  - #service open-iscsi restart
- 查看open-iscsi状态
  - # service open-iscsi status

 Page 42 

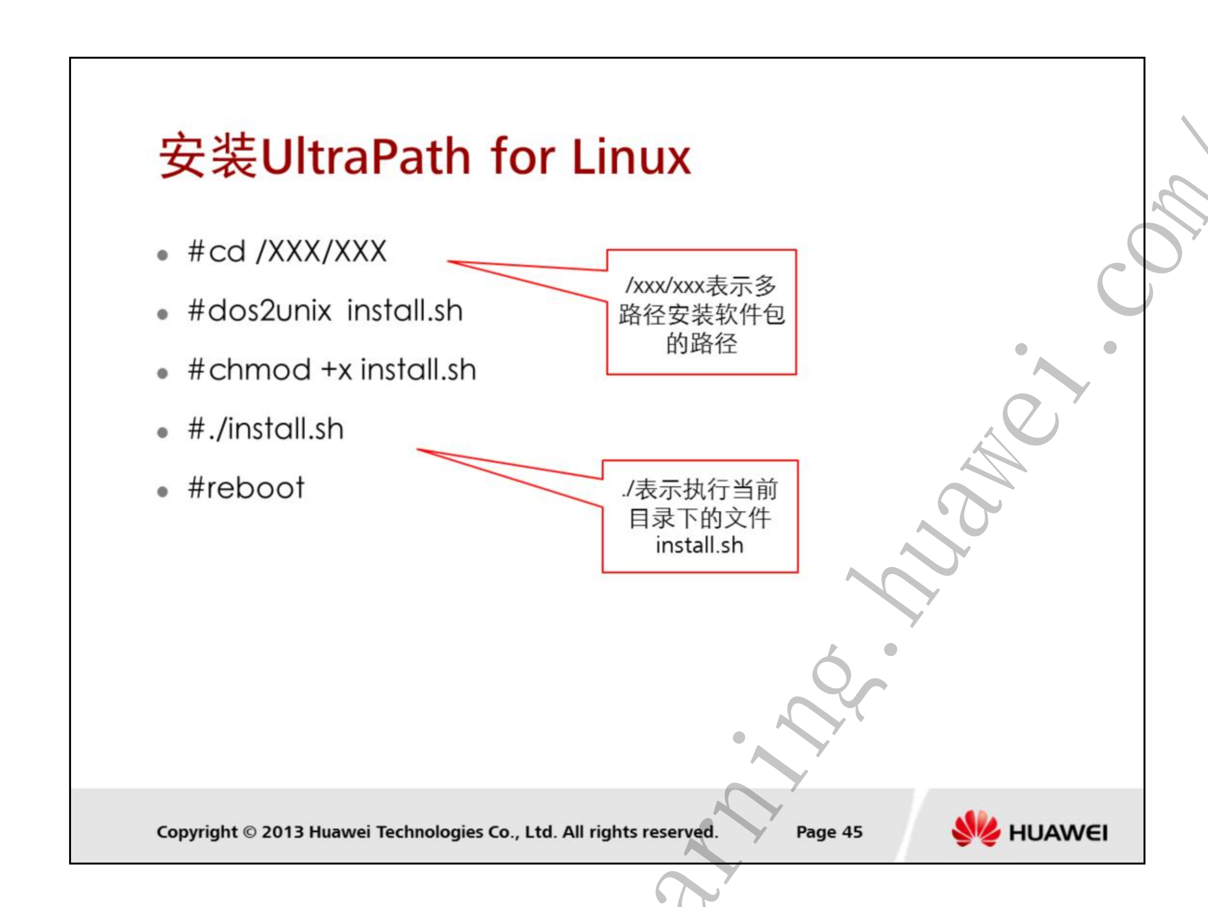
- open-iscsi服务的控制
  - 启动open-iscsi服务
    - /etc/init.d/open-iscsi start (Suse10)
    - /etc/init.d/iscsi start (Redhat5)
  - 停止open-iscsi服务
    - /etc/init.d/open-iscsi stop (Suse10)
    - /etc/init.d/iscsi stop (Redhat5)
  - 查看open-iscsi服务状态
    - /etc/init.d/open-iscsi status (Suse10)
    - /etc/init.d/iscsi status (Redhat5)
  - 重启open-iscsi服务
    - /etc/init.d/open-iscsi restart (Suse10)
    - /etc/init.d/iscsi restart (Redhat5)
  - 配置open-iscsi服务开机时自动启动
    - chkconfig open-iscsi on
- 服务自启动 chkconfig --level 235 aaa on 将服务aaa设置为自动启动



- 当Linux服务器和存储设备连接并登录后，在服务器查看存储设备上分配给服务器的LUN，可以用fdisk –l命令，会看到新的磁盘分区，比如/dev/sdc这样的标识。
- 要使用存储分区，在Linux系统里，还需要对磁盘分区并格式化，然后再挂载,分区可使用fdisk /dev/sdx,(x代表新增的磁盘符），根据命令提示操作，完成分区；然后再格式化建文件系统，使用mkfs.ext3 /dev/sdx(命令x代表新增的磁盘符），最后再挂载分配好的磁盘，可使用如下的命令格式mount /dev/sdx /mnt/sdx。



- 操作步骤：
  - 步骤1 查询应用服务器的Linux 操作系统的版本及位数信息。
  - 步骤2 确认使用于当前Linux 版本的UltraPath for Linux 安装包准备就绪。
  - 步骤3 UltraPath for Linux 的升级需要先完全卸载原来的版本后重新安装新的版本，所以升级时请先中断原来的I/O 应用。
  - 步骤4 如果升级的服务器处于集群环境，请将需要安装UltraPath for Linux 的服务器上的服务转移到集群中其它节点。
  - 步骤5 在安装UltraPath for Linux 软件之前，请确保在应用服务器的本地磁盘上有10MB 的剩余空间。
- 多路径软件兼容性查询方式
  - 在发布的多路径下载软件包中，有对该版本多路径软件兼容性的说明，请仔细检查。



- 步骤1：登录应用服务器。
- 步骤2：运行cd xxx/Linux 命令进入多路径软件安装包目录。

  xxx 表示多路径软件安装光盘的根目录。
- 步骤3：转换install.sh脚本的格式，命令如下：

  #dos2unix ./install.sh
- 步骤4：修改install.sh脚本的执行权限，命令如下：

  #chmod 744 ./install.sh
- 步骤5：执行安装脚本，命令如下：

  #./install.sh
- 步骤6：重启系统，命令如下：

  #reboot

Linux 设备管理

- 内存

  cat /proc/meminfo

  free -m

- 查看CPU信息

  cat /proc/cpuinfo

```
susel:~ # cat /proc/cpuinfo
processor       : 0
vendor_id       : GenuineIntel
cpu family      : 6
model           : 15
model name      : Intel(R) Xeon(R) CPU
stepping        : 8
cpu MHz         : 2325.615
cache size      : 4096 KB
```

- 查看SCSI控制器信息

  lspci | grep SCSI

  cat /proc/scsi/scsi

  cat /proc/scsi/qla2300/1|grep port

- 磁盘

  df –h

  fdisk –l

 Page 46 HUAWEI

- 各版本Linux下查看wwn信息的方法：
- SuSE Linux 9

查看 /proc/scsi/qla2xxx/* ，并以 adapter-port 为关键字过滤即可查看FC HBA卡的WWN信息：

```
# cat /proc/scsi/qla2xxx/* | grep adapter-port
scsi-qla0-adapter-port=21000018822c8a2c;
scsi-qla1-adapter-port=21000018822c8a2d;
```

- SuSE Linux 10

查看 /sys/class/fc_host/host*/port_name 文件的内容即可看到对应FC HBA卡的WWN信息：

```
# cat /sys/class/fc_host/host*/port_name
0x210000e08b907955
0x210000e08b902856
```

- RedHat Linux AS4

查看HBA卡WWN：

```
# grep scsi /proc/scsi/qla2xxx/3
Number of reqs in pending_q= 0, retry_q= 0, done_q= 0, scsi_retry_q= 0
```

Linux 设备管理

- 内存

  cat /proc/meminfo

  free -m

- 查看CPU信息

  cat /proc/cpuinfo

```
susel:~ # cat /proc/cpuinfo
processor       : 0
vendor_id       : GenuineIntel
cpu family      : 6
model           : 15
model name      : Intel(R) Xeon(R) CPU
stepping        : 8
cpu MHz         : 2325.615
cache size      : 4096 KB
```

- 查看SCSI控制器信息

  lspci | grep SCSI

  cat /proc/scsi/scsi

  cat /proc/scsi/qla2300/1|grep port

- 磁盘

  df –h

  fdisk –l

 Page 47 HUAWEI

scsi-qla0-adapter-node=20000018822d7834;

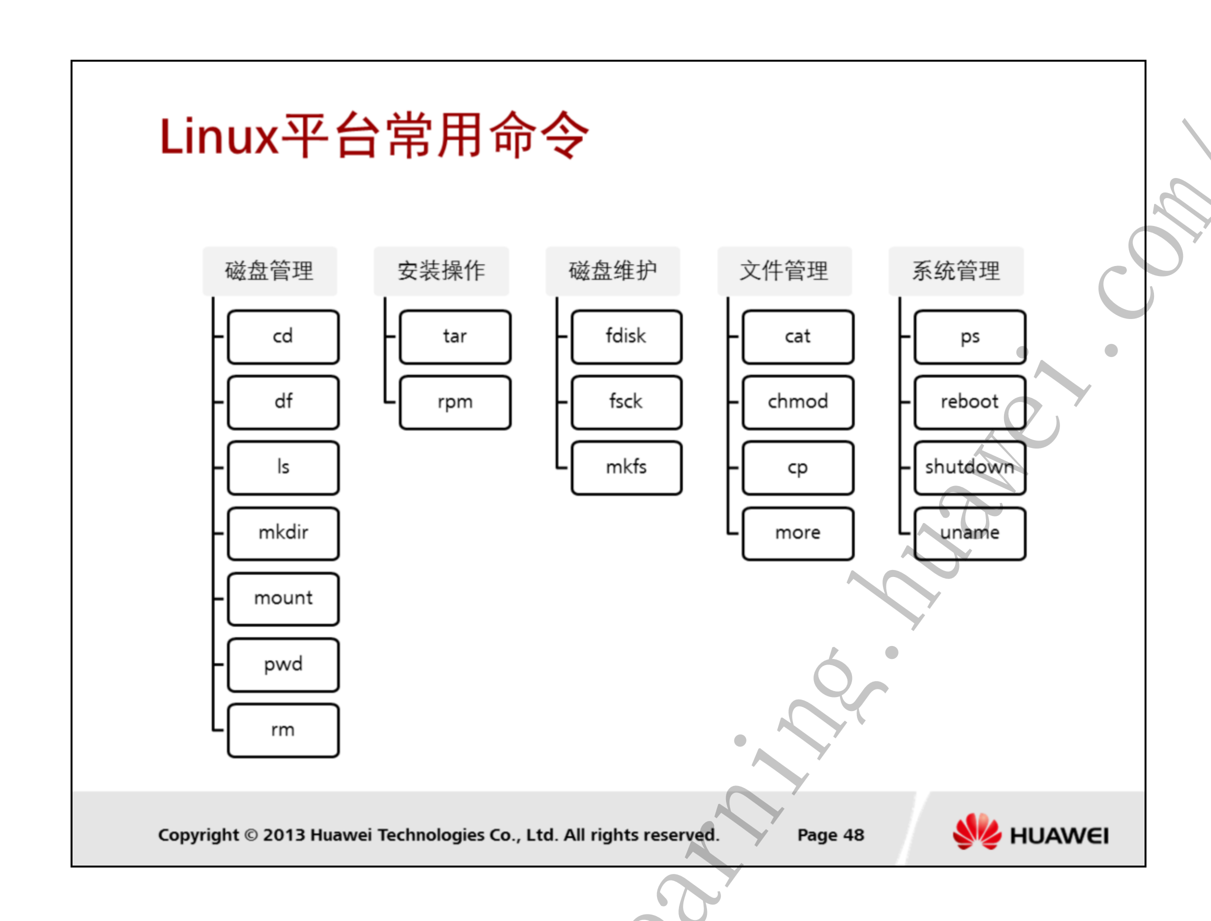
scsi-qla0-adapter-port=21000018822d7834;

scsi-qla0-target-0=202900a0b8423858;

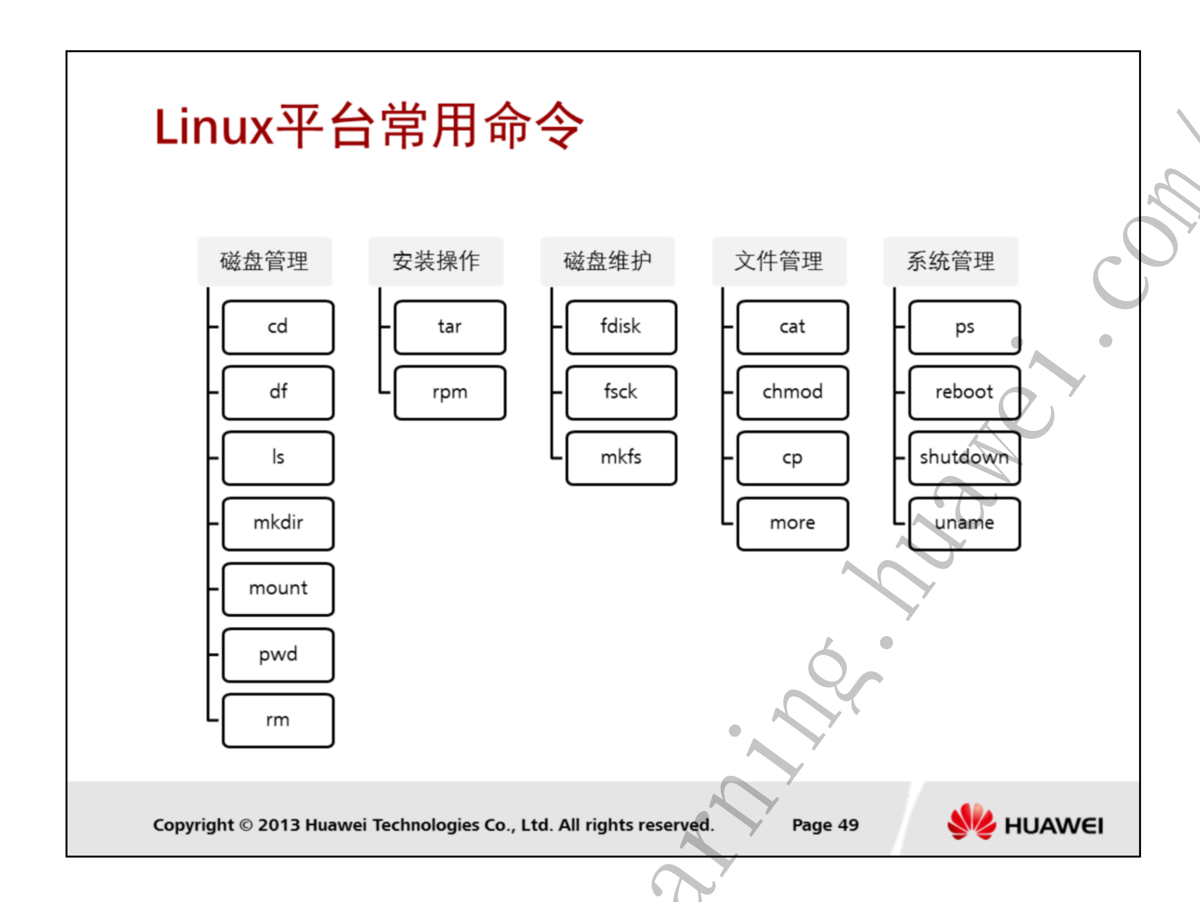
scsi-qla0-port-0=200800a0b8423858:202900a0b8423858:0000e8:1;

- RedHat Linux AS5

  # cat /sys/class/fc_host/hostx/port_name



- **命令　　功能说明：**
- cd　　切换目录
- df　　显示磁盘的相关信息
- ls　　列出目录内容
- mkdir　　建立目录
- mount　　目录挂载
- pwd　　显示工作目录
- rm　　除文件或目录
- tar　　文件归档工具
- rpm　　管理Linux各项套件
- fdisk　　磁盘分区
- fsck　　检查文件系统并尝试修复错误
- mkfs　　建立各种文件系统
- cat　　把档案串连接后传到基本输出
- chmod　　变更文件或目录的权限
- cp　　复制文件或目录



- more 文件分屏显示
- ps 报告程序状况
- reboot 重新开机
- Shutdown 系统关机指令
- uname 显示系统信息
- 具体命令格式及参数请参考命令帮助。



目录



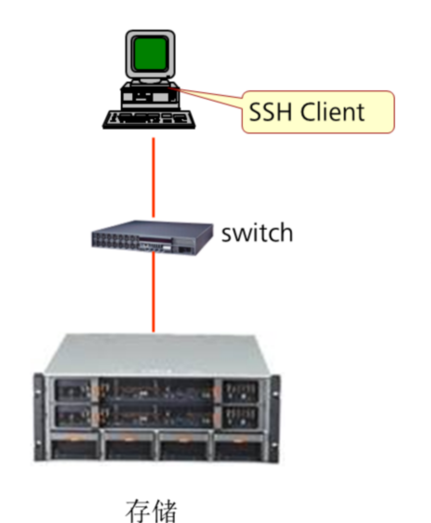
指示灯工作状态总结

- 绿灯亮：模块正常、端口的速率值
- 红灯亮：模块故障
- 红灯闪：模块正在启动、定位端口、定位硬盘
- 绿灯闪：端口正在传输数据、BBU充电、电源模块已接 、硬盘传输数据 电源但未上电、控制器正在启动、端口模块有热拔插请求
- 蓝灯亮：端口的速率值
- 蓝灯闪：端口正在传输数据
- 橙灯亮：端口的速率值
- 橙灯闪：管理网口正在传输数据、BBU正在放电
- 灯不亮：未上电、未连线、端口模块可以拔插、告警灯灭表示正常、1GB的iscsi主机端口的速率灯灭表示速率低于1G

 Page 51 

- 对于S2600、S5000、S5000T端口速率指示灯做了以下总结：
  - 蓝灯亮表示4G的FC主机端口模块速率为4G，8G的FC主机端口模块速率为8G，10G的iscsi主机端口模块速率为10G，miniSAS级联模块与级联框连接速率为6G
  - 绿灯亮表示4G的FC主机端口模块速率为1G或2G，8G的FC主机端口模块速率为4G或2G，10G的iscsi主机端口模块速率为1G，miniSAS级联模块与级联框连接速率为3G
  - 橙灯亮表示1G的iscsi主机端口速率为1G
  - 橙灯灭表示1G的iscsi主机端口速率低于1G

通过CLI界面维护



- 访问方式
  - CLI运行在存储阵列上，通过SSH客户端软件进行访问
  - 同时支持多个客户端进行访问
  - 需要有管理存储阵列的权限
- 功能
  - 根据用户级别对命令进行分类
  - 输入命令，回车查看执行结果
  - 支持通过脚本方式批量处理命令




系统配置导出
OceanStor ISM
系统(S) 性能(P) 事件(E) 帮助(H)
故障统计: 紧急 0 重要 5 警告 1
所有设备(1)
S5500T
配置助手
系统配置
设备信息
存储资源
SAN服务
快照
LUN拷贝
远程复制
一致性组
分裂镜像
映射
主机组
主机
启动器
基本设置
设备时间管理
License管理
事件日志
事件通知设置
事件转储
Trap IP地址管理
导入导出
导入配置文件
导出配置文件
导出运行数据
导出系统日志
性能监控
实时性能监控
查看历史监控数据
性能监控设置
性能告警阈值设置
性能监控文件转储设置
性能监控文件管理
合并监控文件
功耗查看
安全
用户管理
域认证服务器设置
可访问IP地址管理
高级
配置iSNS服务
修改iSCSI设备名称
软件升级
补丁管理
误码统计
系统节能信息
Copyright © 2013 Huawei Technologies Co., Ltd. All rights reserved.
Page 53
HUAWEI

实时性能监控

实时性能监控

本功能能够帮助您对所管理设备的各项性能进行实时监控，同时支持统计的统计项不能超过16个。进行监控前，请确认性能监控功能是启用状态。

设备序列号：210235G6GTZ0BB000003　设备名称：S5500T　采样周期(秒)：5

统计对象

对象类型：控制器

- [x] A
- [x] B

统计项

统计类型：基本统计项

- [x] 吞吐量（IOPS）
- [ ] 带宽
- [ ] 内存占用率
- [x] 平均I/O响应时间
- [ ] 脏页比例
- [x] Cache读利用率
- [ ] Cache写利用率

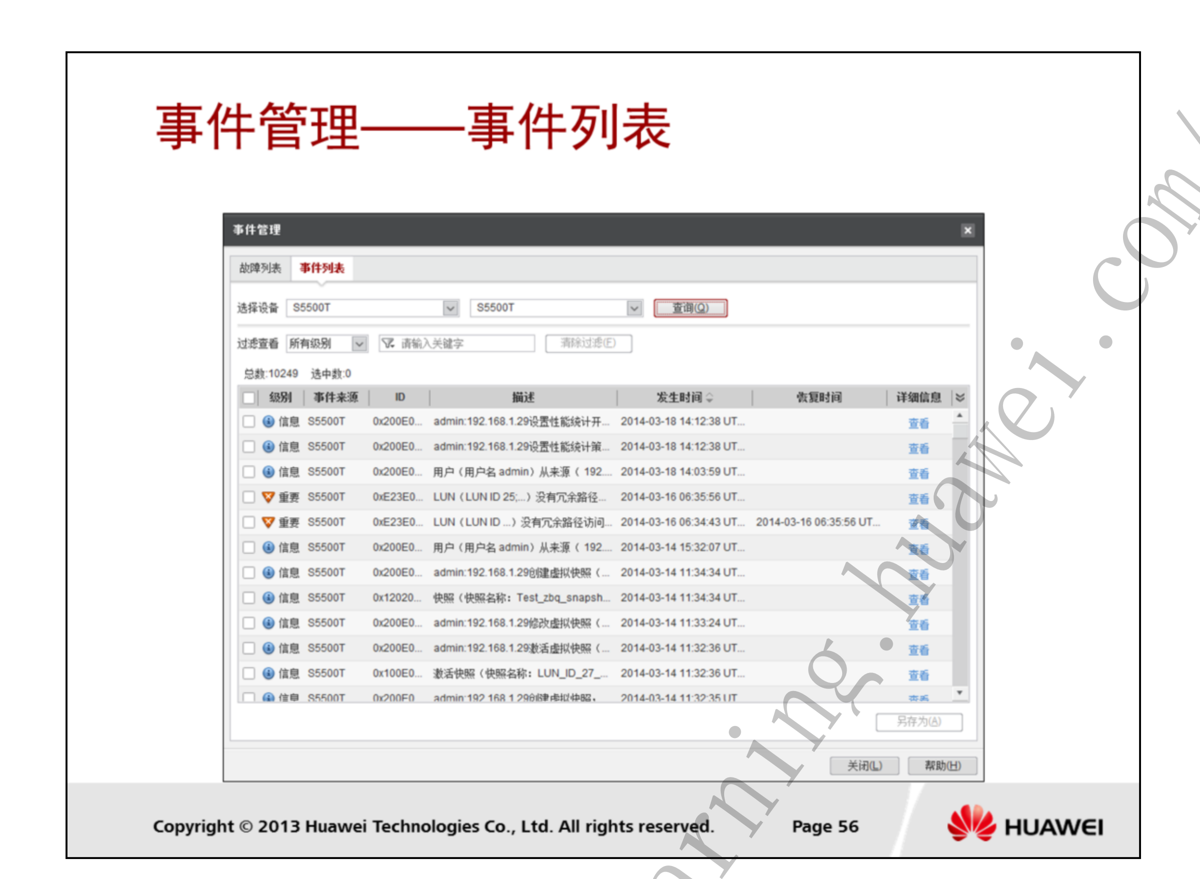
总数:6　选中数:0

图例	对象类型	对象名称	比例	统计项	单位
	控制器	A	自适应 (1:1)	吞吐量（IOPS）	IO/s
	控制器	A	自适应 (1:1)	平均I/O响应时间	ms
	控制器	A	自适应 (1:1)	Cache读利用率	%
	控制器	B	自适应 (1:1)	吞吐量（IOPS）	IO/s
	控制器	B	自适应 (1:1)	平均I/O响应时间	ms

关闭(L)　帮助(H)

HUAWEI


事件管理——故障列表
事件管理
故障列表
事件列表
过滤查看
所有设备
所有事件来源
所有级别
清除过滤(F)
总数:6
选中数:0
级别 | 事件来源 | ID | 描述 | 发生时间 | 详细信息
重要 | S5500T | 0xE23E0001 | LUN（LUN ID 25;...）没有冗余路径访问。 | 2014-03-16 06:35:56 UTC+08:00 | 查看
警告 | S5500T | 0xE01F90008 | 系统中缺少容量大于或等于278GB的SAS热备... | 2014-03-09 18:31:01 UTC+08:00 | 查看
重要 | S5500T | 0xE23E0004 | 主机（主机名 N8500_01）上未安装UltraPath... | 2014-02-25 16:23:49 UTC+08:00 | 查看
重要 | S5500T | 0xE23E0004 | 主机（主机名 N8500_02）上未安装UltraPath... | 2014-02-25 16:20:12 UTC+08:00 | 查看
重要 | S5500T | 0xE23E0004 | 主机（主机名 N8500）上未安装UltraPath多路... | 2014-02-22 17:23:11 UTC+08:00 | 查看
重要 | S5500T | 0xE23E0004 | 主机（主机名 VTL_1）上未安装UltraPath多路... | 2014-02-21 11:28:33 UTC+08:00 | 查看
刷新(R)
关闭(L)
帮助(H)
Copyright © 2013 Huawei Technologies Co., Ltd. All rights reserved.
Page 55
HUAWEI


事件管理——事件列表
事件管理
故障列表
事件列表
选择设备 S5500T S5500T 查询(Q)
过滤查看 所有级别 请输入关键字 清除过滤(E)
总数:10249 选中数:0
级别 事件来源 ID 描述 发生时间 恢复时间 详细信息
信息 S5500T 0x200E0... admin:192.168.1.29设置性能统计开... 2014-03-18 14:12:38 UT... 查看
信息 S5500T 0x200E0... admin:192.168.1.29设置性能统计策... 2014-03-18 14:12:38 UT... 查看
信息 S5500T 0x200E0... 用户（用户名 admin）从来源（ 192.... 2014-03-18 14:03:59 UT... 查看
重要 S5500T 0xE23E0... LUN（LUN ID 25;...）没有冗余路径... 2014-03-16 06:35:56 UT... 查看
重要 S5500T 0xE23E0... LUN（LUN ID ...）没有冗余路径访问... 2014-03-16 06:34:43 UT... 2014-03-16 06:35:56 UT... 查看
信息 S5500T 0x200E0... 用户（用户名 admin）从来源（ 192.... 2014-03-14 15:32:07 UT... 查看
信息 S5500T 0x200E0... admin:192.168.1.29创建虚拟快照（... 2014-03-14 11:34:34 UT... 查看
信息 S5500T 0x12020... 快照（快照名称：Test_zbq_snapsh... 2014-03-14 11:34:34 UT... 查看
信息 S5500T 0x200E0... admin:192.168.1.29修改虚拟快照（... 2014-03-14 11:33:24 UT... 查看
信息 S5500T 0x200E0... admin:192.168.1.29激活虚拟快照（... 2014-03-14 11:32:36 UT... 查看
信息 S5500T 0x100E0... 激活快照（快照名称：LUN_ID_27_... 2014-03-14 11:32:36 UT... 查看
另存为(A)
关闭(L)
帮助(H)
Copyright © 2013 Huawei Technologies Co., Ltd. All rights reserved.
Page 56
HUAWEI

软件升级

- 网管下发一次升级命令，系统自动判断需要升级的部件，自动升级所有需要升级的部件，升级过程中，网管能实时监控升级进程的进行，实时反馈升级的结果。



Page 57

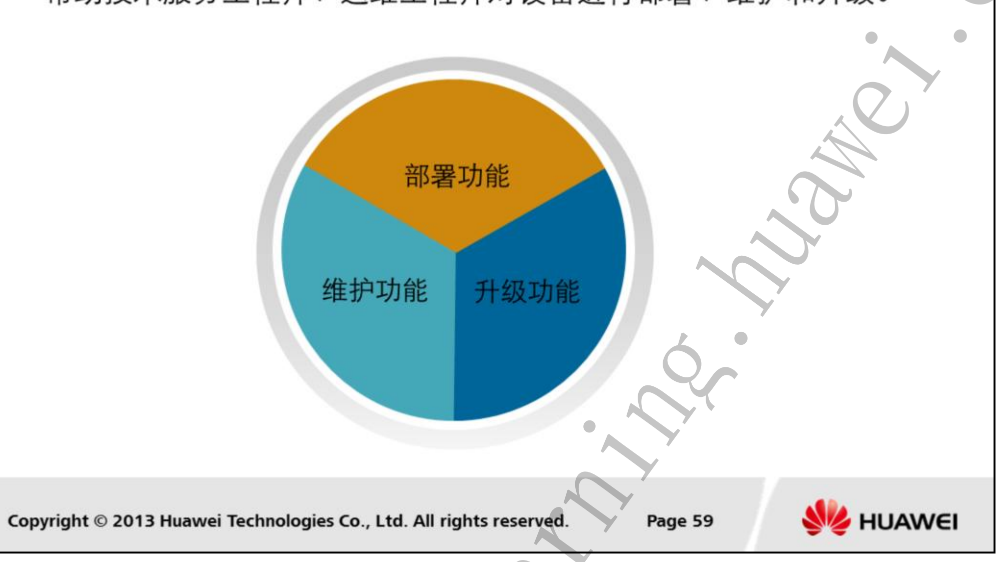
HUAWEI

- 当前升级有两种方式：在线升级和离线升级。
- 在线升级：其是一种不中断业务的升级方式，有严格的升级前检查条件，只有所有条件通过之后才能选择在线升级。
- 离线升级：其是一种必须先停主机业务的升级方式，升级时间比在线升级要短。
- 升级主要流程：
  - 解压升级包
  - 升级管理芯片(SES、BMC)
  - 升级备用电池单元（BBU）
  - 升级接口卡固件（SAS、TOE、FCoE）
  - 升级逻辑器件（CPLD、FGPA）
  - 升级系统软件
  - 系统重启
  - 升级BIOS
  - 版本校验



目录





- 部署功能
    - 部署功能可以帮助技术服务工程师、运维工程师进行设备初始化。目前部署工具已经实现工具独立安装，实现工具独立安装之后，部署工具可以通过工具Store进行安装，卸载或升级。
- 维护功能
    - 维护功能可以帮助技术服务工程师、运维工程师对设备进行日常检查、信息收集等操作。
- 升级功能
    - 升级功能可以帮助技术服务工程师、运维工程师对设备进行升级。

运行环境

- 运行前提
  - 已安装JRE 1.6.0_20以上或JRE 1.7.0以下版本

配置项	版本信息
操作系统	阵列初始化部署、VIS系统部署、S8000系统部署、集群NAS部署、统一存储部署只支持32位的Windows XP、Windows Server 2003和Windows 7。 其它工具支持32位和64位的的Windows XP、Windows Server 2003和Windows 7。
Java运行环境	JRE 1.6.0_20以上或JRE 1.7.0以下版本





- 1、功能导航栏 显示ToolKit的所有功能模块。
- 2、语言切换栏 切换不同语言。
- 3、帮助、关于和系统设置按钮
- 4、子功能入口
- 5、当前位置 显示当前所在的位置。

部署



- 在部署项里，主要包括以下设备的部署：
- 阵列批量部署
- CSS系统部署
- 集群NAS存储系统部署



- 在默认安装时，维护里面会安装两个子功能部件，“巡检及续保检查”和“信息收集”，在工具store选项里，还有其它的维护工具可下载安装。



- 在升级选项里，包括以下子功能部件：
- 设备升级
- 跨版本升级
- 集群NAS升级
- 补丁工具



总结

- 存储初始化配置流程
- 存储端基础配置
- 主机端基础配置
- 存储运维管理





思考题

1. 应用服务器侧要安装哪些软件才能正常使用存储设备映射的LUN?
2. Windows平台和linux平台iSCSi的配置有什么不同?
3. UltraPath 软件的主要功能有哪些?
4. UltraPath安装有哪些注意事项?
5. 存储设备上创建LUN时需要考虑文件系统的哪些参数?
6. linux下iscsi initiator软件的常用命令?





练习题

多选题

1、多路径软件安装的注意事项有（　　）

A. 检查UltraPath的兼容性

B. 关闭HBA卡自带的Failover和Failback功能

C. 服务器安装相同HBA卡

D. 存储系统正确配置端口模块

判断题

1、在windows环境下，initiator配置添加目标器的时候，目标器的IP地址配置为存储设备的管理IP。（T　or　F）



- 习题答案
  - 多选题

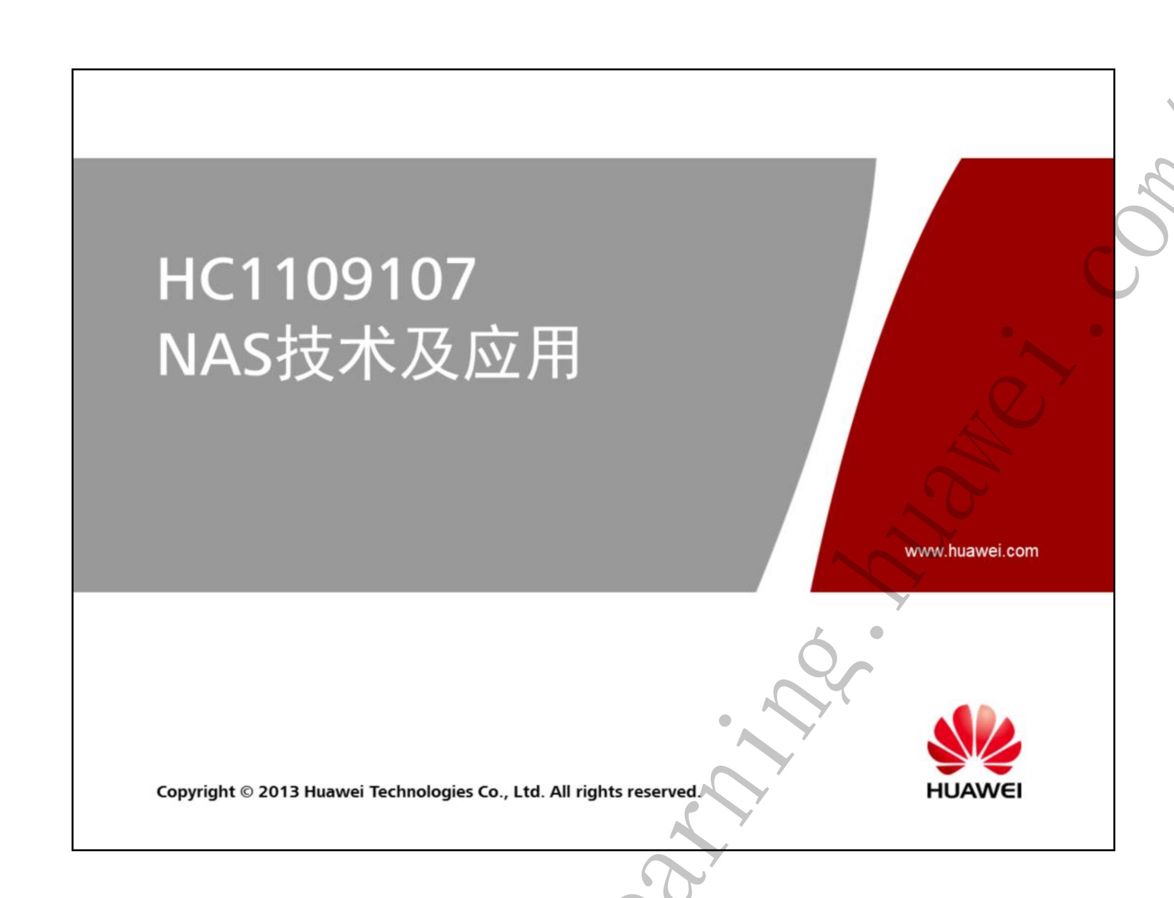
  1、ABCD
  - 单选题

  1、F


Thank you
www.huawei.com
Copyright © 2013 Huawei Technologies Co., Ltd. All rights reserved.
Page 68
HUAWEI

HC1109107 NAS 技术及应用




HC1109107
NAS技术及应用
www.huawei.com
Copyright © 2013 Huawei Technologies Co., Ltd. All rights reserved.
HUAWEI



目标

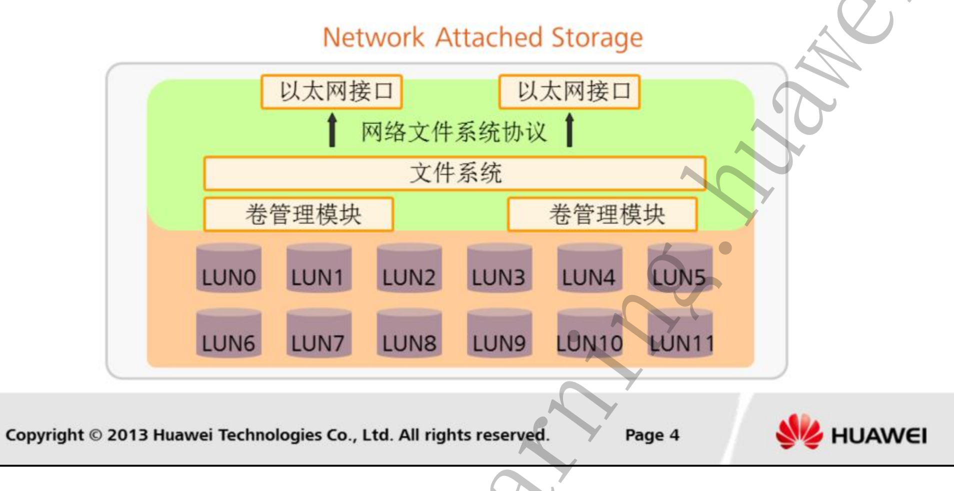
- 学习完本课程后，您将能够：
  - 了解NAS组成结构及实现
  - 掌握NAS文件共享协议NFS以及CIFS
  - 了解NAS系统文件IO与性能
  - 了解SAN与NAS的区别及联系
  - 了解华为NAS产品





目录

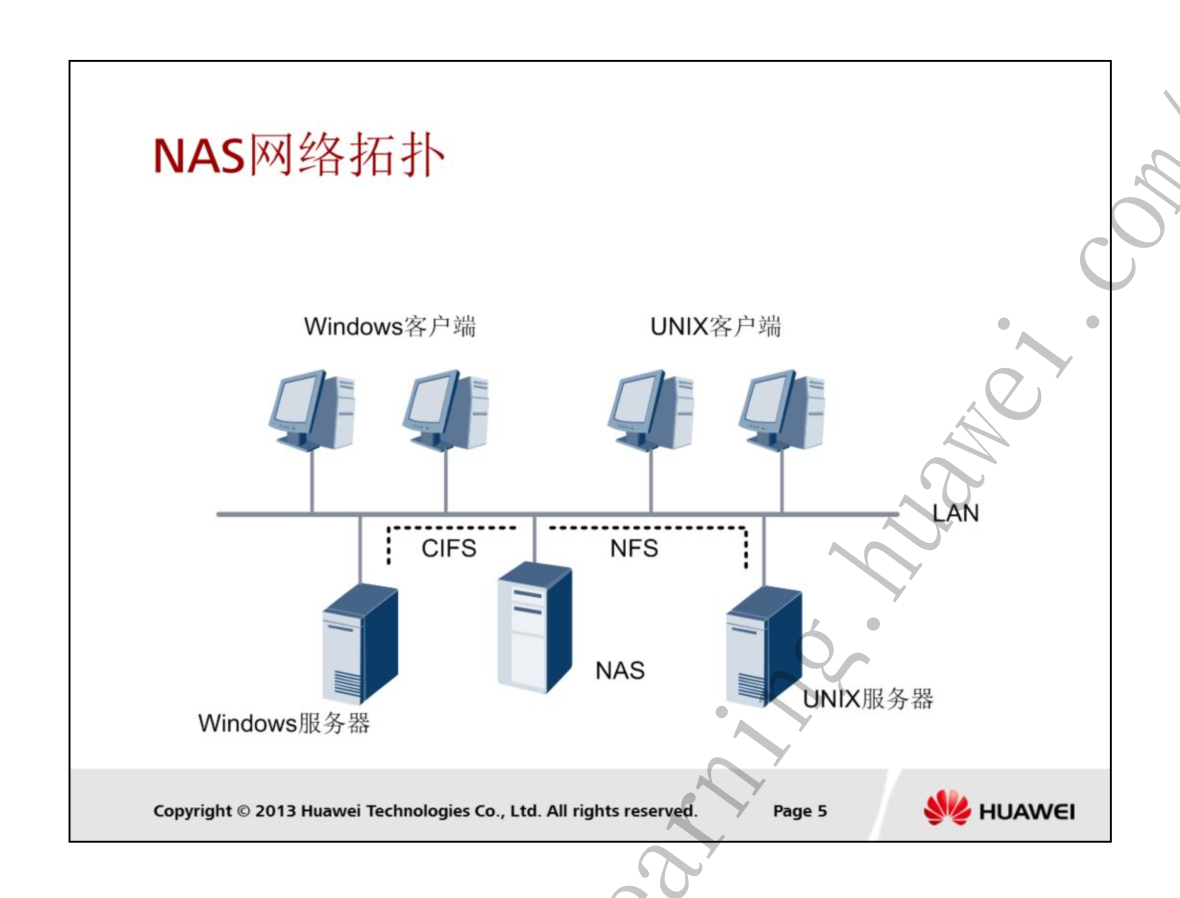




NAS和SAN最大的区别就在于NAS有文件操作和管理系统，而SAN却没有这样的系统功能，其功能仅仅停留在文件管理的下一层，即数据管理。SAN和NAS并不是相互冲突的，是可以共存于一个系统网络中的，但NAS通过一个公共的接口实现空间的管理和资源共享，SAN仅仅是为服务器存储数据提供一个专门的快速后方存储通道。

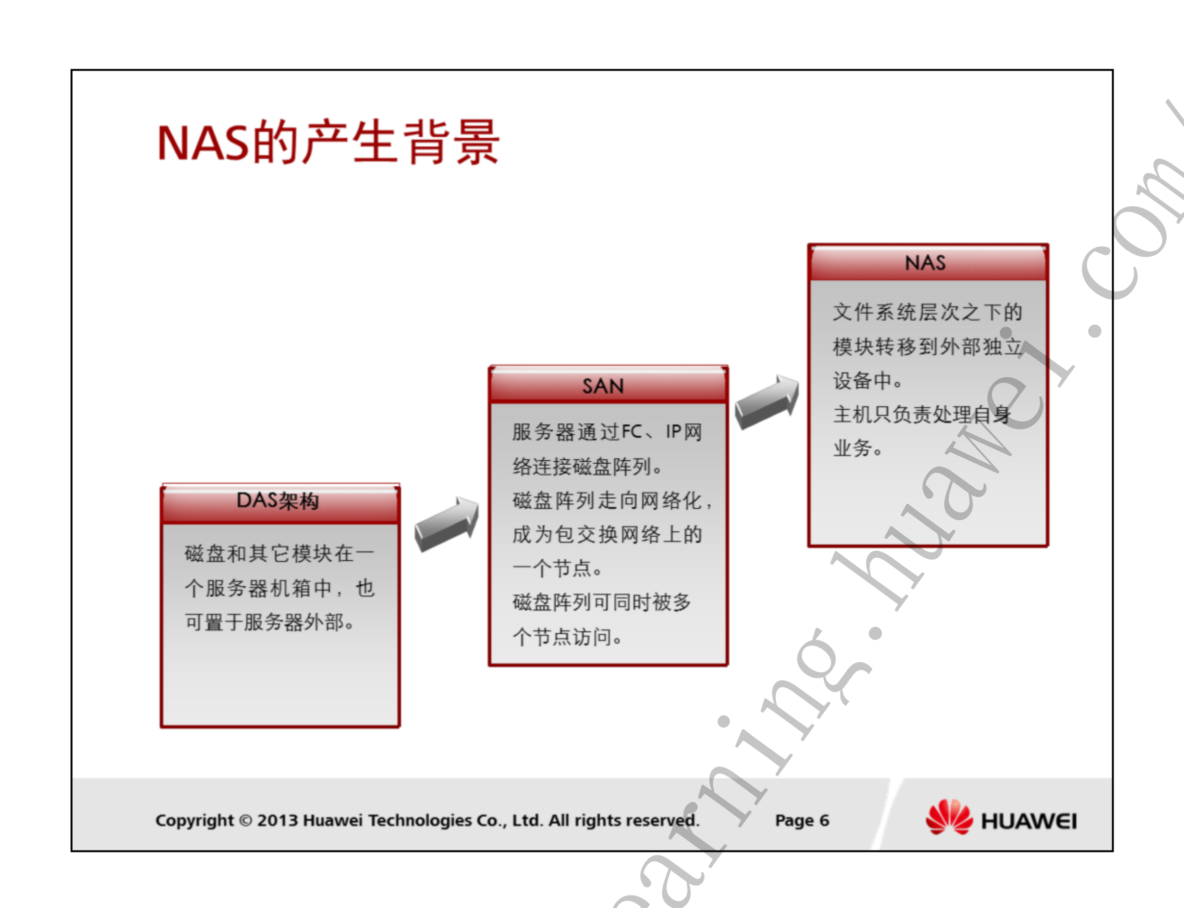
为什么FTP文件服务不属于NAS?

FTP只能将文件传输到本地的目录之后才能执行，而网络文件系统可以允许直接问原端的文件，不需要将数据复制到本地再访问。



NAS可作为网络节点，直接接入网络中，理论上NAS可支持各种网络技术，支持多种网络拓扑，但是以太网是目前最普遍的一种网络连接方式，我们主要讨论是以以太网为网络基础的NAS环境。

NAS本身能够支持多种协议（如NFS、CIFS、FTP、HTTP等），而且能够支持各种操作系统。通过任何一台工作站，采用IE或Netscape浏览器就可以对NAS设备进行直观方便的管理。



传统的DAS存储系统尽管使用方便，但这种模式是直接将存储设备连接到服务器上的。一方面，当存储容量增加时，这种方式很难扩展;另一方面，当服务器出现异常时，会使数据不可获得，容错性差;再者，存在着存储数据无法被其他服务器共享，扩充容量则需要关闭整个系统，远程管理不方便等诸多缺陷。于是便有了网络存储技术的出现，网络存储技术分为两类，SAN和NAS,在SAN环境中，存储设备通过网络与服务器相连，可以有更多的的服务器访问存储设备提供的存储资源，存储设备提供数据块级别的服务。NAS是一种向用户提供文件级服务的专用数据存储设备，直接连到网络上，不再挂接服务器后端，避免给服务器增加I/O负载,服务器只负责处理自身业务。

集群NAS存储系统概念与特点

- 集群技术
  - 定义：
    - 集群是由一组相互独立的服务器组成，对外表现为单一服务器，提供高可靠性服务。
  - 特点：
    - 统一命名
    - 高可靠性
    - 性能扩展
    - 共享数据空间

集群系统

增加组件

业务网络

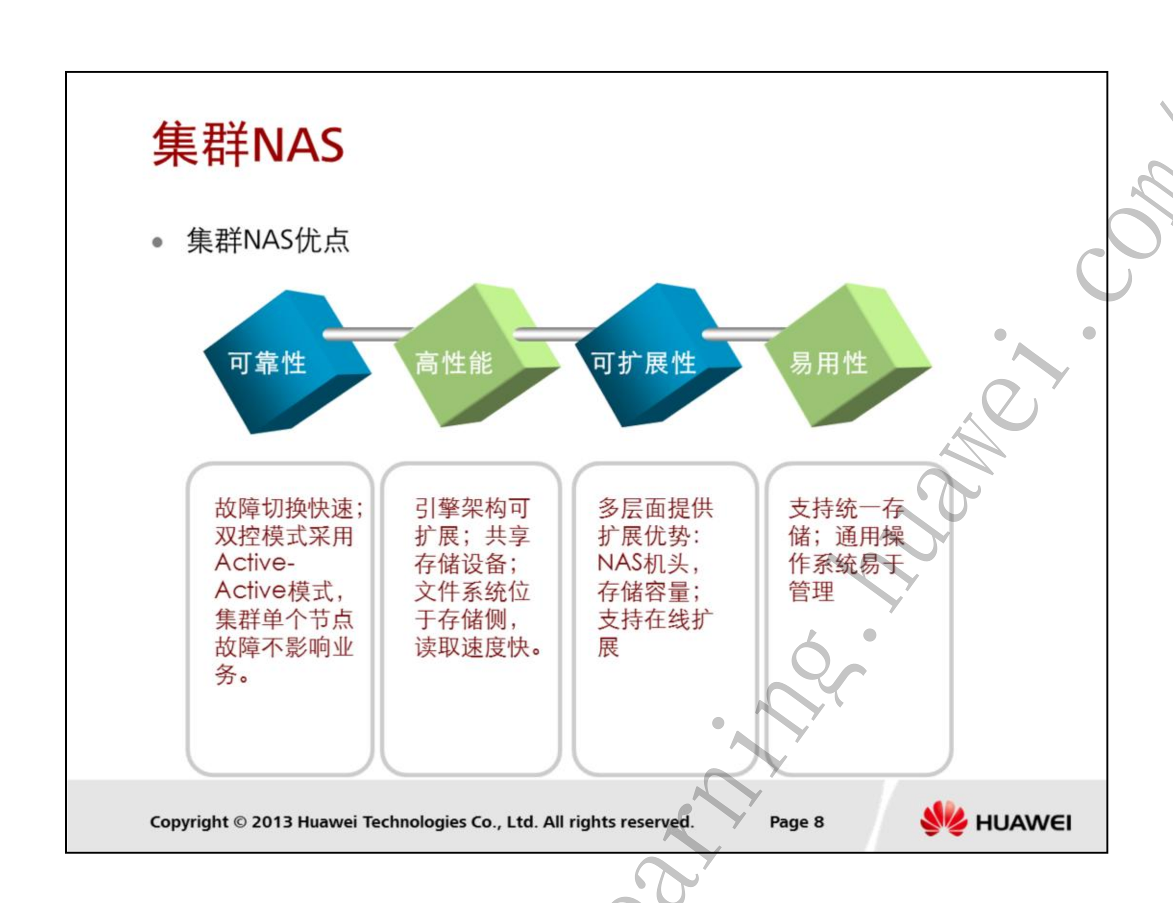
Page 7

HUAWEI

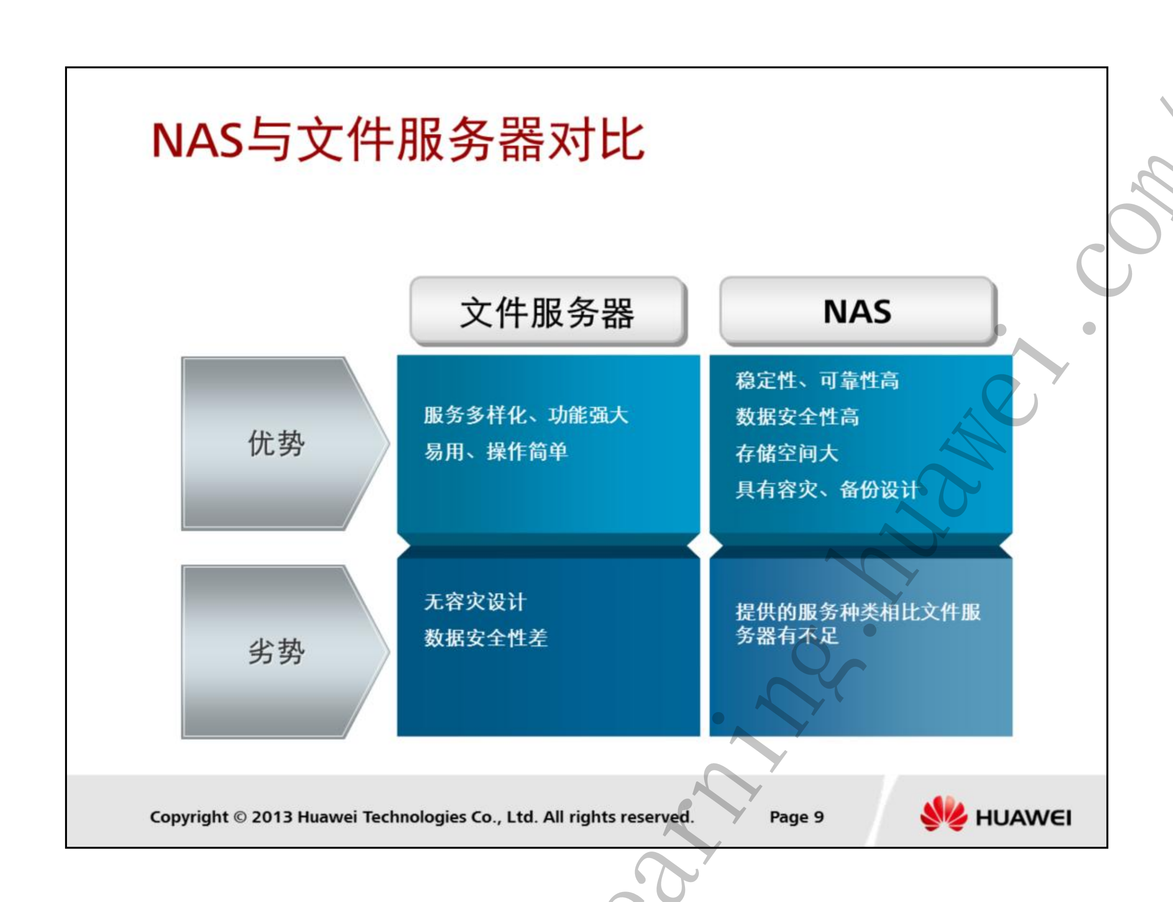
- 集群技术概念:

　　一组相互独立的服务器在网络中表现为单一的系统，并以单一系统的模式加以管理。此单一系统为客户工作站提供高可靠性的服务。

- 集群技术特点：
  - 统一命名：大多数模式下，集群中所有的计算机拥有一个共同的名称，集群内任一系统上运行的服务可被所有的网络客户所使用。
  - 高可靠性：集群必须可以协调管理各分离的组件的错误和失败，集群内单一系统的失败由其他集群成员来弥补，对客户是不可见的。集群内部各节点服务器通过一内部局域网相互通讯。当一台节点服务器发生故障时，这台服务器上所运行的应用程序将在另一节点服务器上被自动接管。当一个应用服务发生故障时，应用服务将被重新启动或被另一台服务器接管。当以上的任一故障发生时，客户都将能很快连接到新的应用服务上。
  - 性能扩展：并可透明地向集群中加入组件，提升系统性能。
  - 共享数据空间：一个集群包含多台(至少二台)拥有共享数据存储空间的服务器。任何一台服务器运行一个应用时，应用数据被存储在共享的数据空间内。每台服务器的操作系统和应用程序文件存储在其各自的本地储存空间上。



集群NAS相较于传统NAS，由于引擎采用了集群架构，带来了多方面的优势，引擎集群之间采用双活模式，可靠性更高。集群架构易扩展，增加引擎能线性提高性能；易扩展，新的引擎可直接加入集群，在线扩容，对业务不造成影响。易管理，对集群内的引擎节点可统一管理。



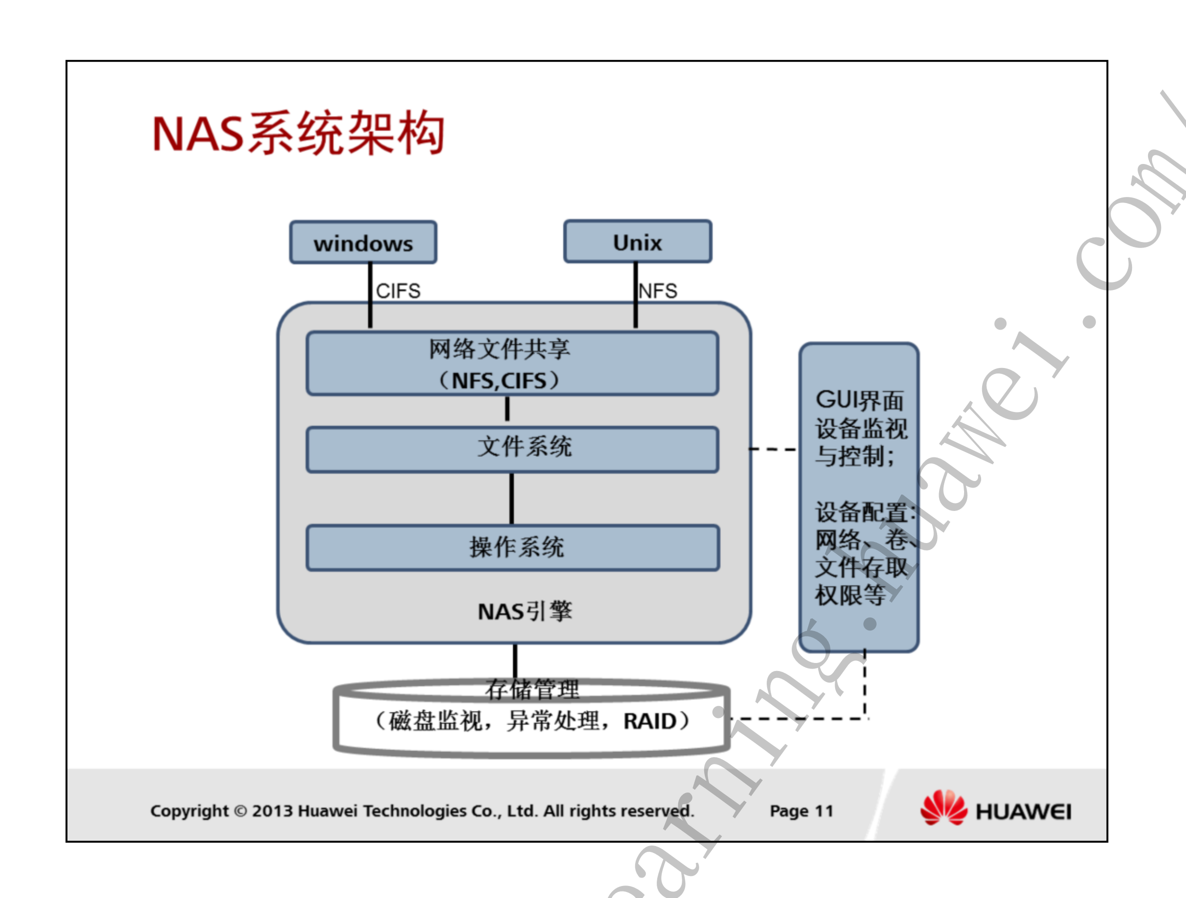
文件服务器主要任务是为网络上的计算机提供多样化的服务，如在文件共享及处理、网页发布、FTP、电子邮件服务等方面有明显的优势，这主要得益于文件服务备器通常是采取高性能的CPU，与NAS相比它在数据备份、数据安全等方面并不占优势。

通过对比不难看出，文件服务器相对于NAS综合功能特别是在文件处理能力方面更为强大，但文件服务器在数据的备份和恢复却远没有NAS的功能完善，系统稳定性亦没NAS的好，存储容量空间没有NAS大易扩展，同时在数据安全如数据容灾方面NAS更具优势。在这种状况下，两者无法相互替代。



目录





NAS系统软件设计的基本要求是较高的稳定性和I/O吞吐率，并能满足数据共享、数据备份、安全配置、设备管理等要求。该结构分为五个模块：操作系统、存储管理器、文件系统、网络文件共享和GUI管理模块。

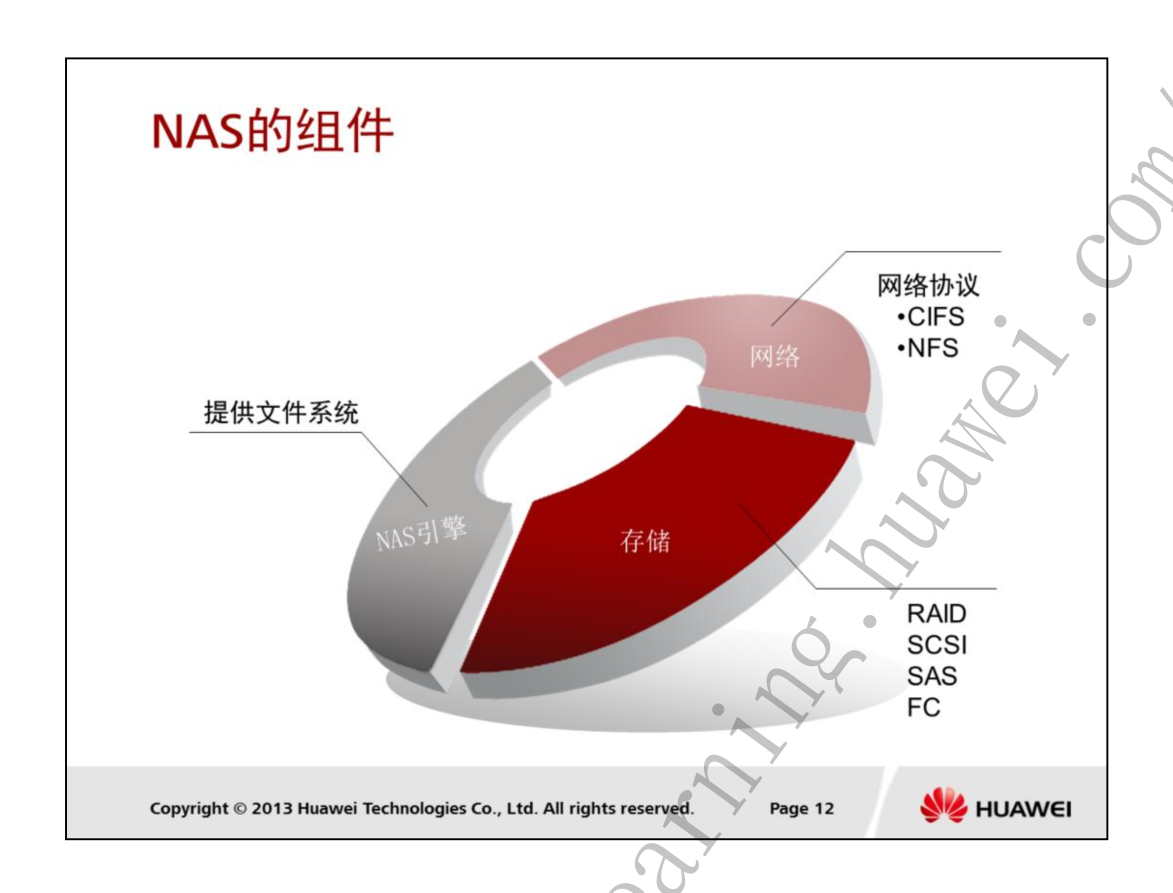
鉴于Linux、FreeBSD等免费的开放源码操作系统具有稳定、可靠、高效的优秀特性，现在大部分NAS设备都是基于此类操作系统开发的。

存储管理器的主要功能是磁盘和分区的管理，主要包括磁盘的监测与异常处理和逻辑卷的配置管理，一般应支持磁盘的热插拔、热替换等功能和RAID0、RAID1、RAID5类型的逻辑卷。存储管理器实现简化的、集中的存储管理功能，保证数据的完整性，并增强数据的可用性。

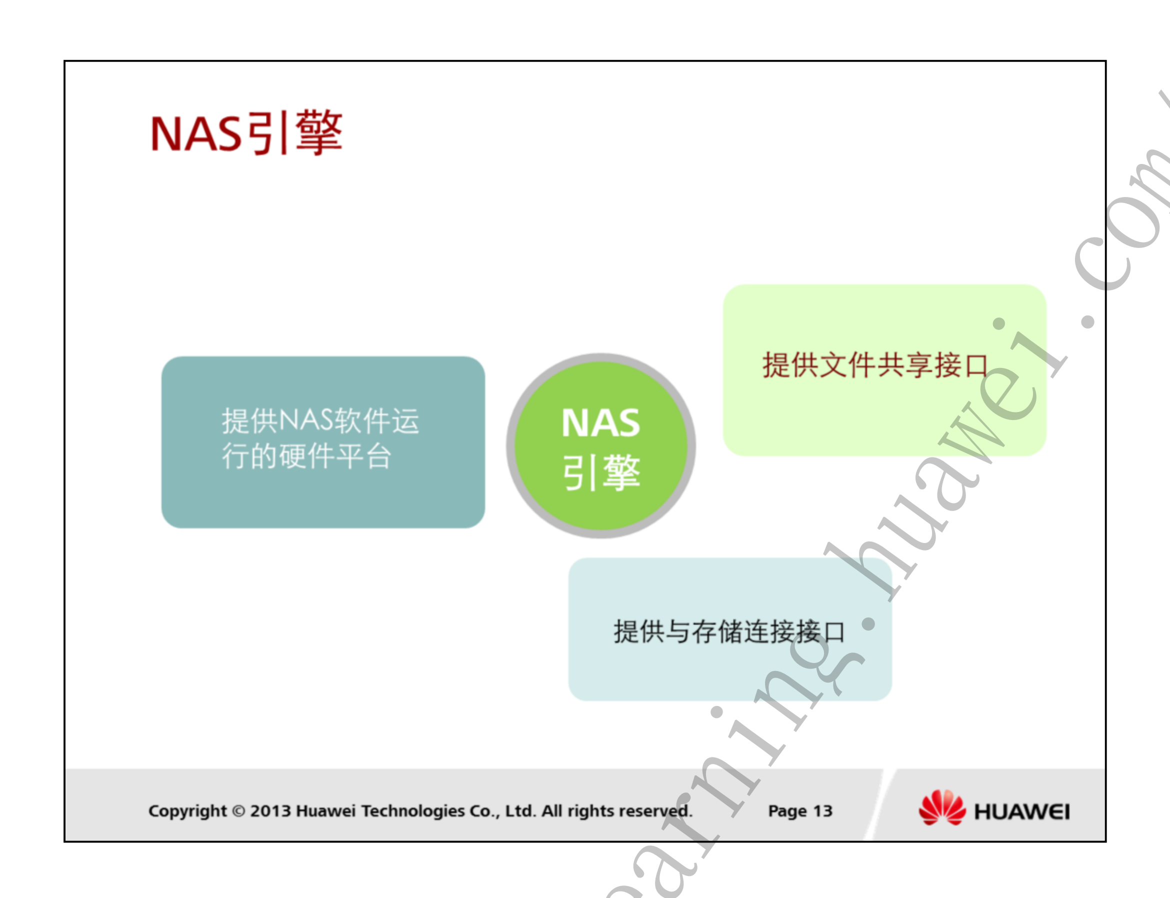
文件系统提供持久性存储和管理数据的手段，它必须是32位或以上并能支持多用户，应具备日志文件系统功能以使系统在崩溃或掉电重启后能迅速恢复文件系统的一般性和完整性，进一步提供NAS的可用性。此外，文件系统还应具有快照功能。快照不仅能恢复被用户错误修改或删除的文件，而且能实现备份窗口为零的文件系统活备份。

网络文件共享一般支持以下一些文件传输和共享协议，如FTP和HTTP协议、Unix系统的NFS、Windows系统的CIFS、Novell系统的NCP（Novell core protocol）、Apple系统的AFP（appletalk file protocol）等，因此NAS设备具有较好的协议独立性。

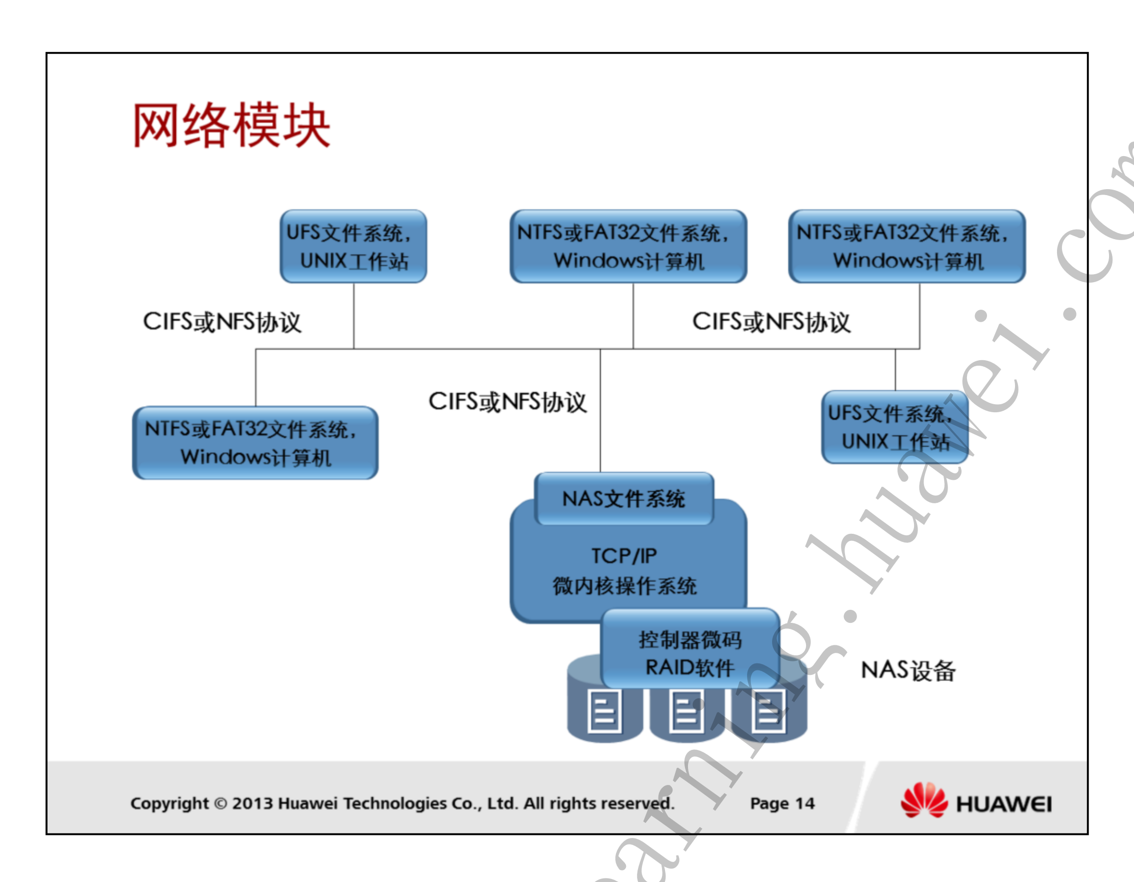
GUI管理提供给系统管理员一个友好的界面，使之仅通过web浏览器操作就能远程监视和管理NAS设备的系统参数，如：网络配置、用户与组管理、卷以及文件共享权限等。



- 存储部分功能模块提供了真正的物理存储空间，主要技术是RAID、SCSI、SAS、FC等技术。
- 控制器部分指NAS引擎部分，这部分提供了NAS底层所使用的文件系统，以及承载文件系统、各种前端协议的操作系统。
- 网络部分提供了和用户交互的网络协议，主要包括NFS和CIFS，用户最终通过这些协议访问存储空间。



NAS引擎是NAS集群软件运行的硬件平台，通过软件的处理，将后端存储提供的卷以NAS方式提供给客户使用，需要较好的I/O处理能力、网络带宽和可靠性



NAS设备中所包含的标准文件系统可以对公用互联网文件系统（CIFS）或是网络文件系统（NFS）提供支持，也有可能同时支持二者。在许多情况下，它都使用标准的网络文件系统来作为NAS专用文件系统的接口。大多数NAS设备需要用这种方式来管理其自身的存储资源。

NFS（网络文件系统）是Unix系统间实现磁盘文件共享的一种方法，支持应用程序在客户端通过网络存取位于服务器磁盘中数据的一种文件系统协议。其实它包括许多种协议，最简单的网络文件系统是网络逻辑磁盘，即客户端的文件系统通过网络操作位于远端的逻辑磁盘。现一般在Unix主机之间采用Sun开发的NFS（Sun），它能够在所有Unix系统之间实现文件数据的互访，逐渐成为主机间共享资源的一个标准。

CIFS是由微软开发的，用于连接Windows客户机和服务器。经过Unix服务器厂商的重新开发后，它可以用于连接Windows客户机和Unix服务器，执行文件共享和打印等任务。它最早的由来是NetBIOS，这是微软开发的在局域网内实现基于Windows名称资源共享的API。之后，产生了基于NetBIOS的NetBEUI协议和NBT(NetBIOS OVER TCP/IP)协议。NBT协议进一步发展为SMB（Server Message Block Potocol）和CIFS（Common Internet File System，通用互联网文件系统）协议。其中，CIFS用于Windows系统，而SMB广泛用于Unix和Linux，两者可以互通。SMB协议还被称作LanManager协议。CIFS支持与SMB的服务器通信而实现共享。微软操作系统家族和几乎所有Unix服务器都支持SMB协议/SAMBA软件包。

存储

- 存储为NAS系统提供了真正的物理存储空间，数据通过NAS引擎处理以后，将数据保存到存储设备中。







目录

1. NAS的产生与发展
2. NAS系统组成与部件
3. **NAS文件共享协议CIFS,NFS**
4. NAS文件系统IO与性能
5. SAN与NAS比较
6. 华为实现与应用



什么是CIFS?

- CIFS (Common Internet File System)，通用Internet文件系统 ，一个新提出的协议，它使程序可以访问远程Internet计算机上的文件并要求此计算机的服务。
- 架构
  - 客户/服务器模式
- 应用
  - Windows系统共享文件的环境
- 传输协议
  - TCP/IP

Page 17



CIFS使用客户/服务器模式，客户程序请求远在服务器上的服务器程序为它提供服务，服务器获得请求并返回响应，用在Windows系统共享文件的环境。当NAS系统对Windows系统服务器提供存储资源共享时，采用CIFS文件系统。

什么是NFS？

- NFS（Network File System）–网络文件系统
  - NFS（网络文件系统）是Unix系统间实现磁盘文件共享的一种方法，支持应用程序在客户端通过网络存取位于服务器磁盘中数据的一种文件系统协议。
- 架构
  - 客户端/服务器架构
- 应用
  - 主要应用在UNIX环境

 Page 18



NFS (Network File System，网络文件系统)是当前主流异构平台共享文件系统之一。主要应用在UNIX环境下。 最早是由SUN microsystem开发，现在能够支持在不同类型的系统之间通过网络进行文件共享，广泛应用在FreeBSD、SCO、Solaris等等异构操作系统平台， 允许一个系统在网络上与它人共享目录和文件。通过使用NFS，用户和程序可以象访问本地文件一样访问远端系统上的文件，使得每个计算机的节点能够像使用本地资源一样方便地使用网上资源。换言之，NFS 可用于不同类型计算机、操作系统、网络架构和传输协议运行环境中的网络文件远程访问和共享。

NFS的工作原理是使用客户端/服务器架构，由一个客户端程序和服务器程序组成。服务器程序向其它计算机提供对文件系统的访问，其过程就叫做“输出”。NFS 客户端程序对共享文件系统进行访问时，把它们从 NFS 服务器中“输送”出来。文件通常以“块”为单位进行传输. 其尺寸是 8K (虽然它可能会将操作分成更小尺寸的分片).NFS 传输协议用于服务器和客户机之间文件访问和共享的通信，从而使客户机远程地访问保存在存储设备上的数据。

CIFS和NFS对比

- 如果文件系统已经设置为CIFS共享，则此文件系统只能设置为只读的NFS共享。
- 如果文件系统已经设置为NFS共享，则此文件系统只能设置为只读的CIFS共享。

协议	传输协议	客户端要求	故障影响	效率	支持操作系统
CIFS	TCP/IP	操作系统集成，无需添加额外软件	大	高	Windows
NFS	TCP或UDP	需要额外软件	小，可自恢复交互过程	低	Unix

Page 19



1、CIFS面向网络连接的共享协议，对网络传输的可靠性要求高，常使用TCP/IP；NFS是独立于传输的，可使用TCP或UDP；

2、NFS缺点之一，是要求client必须安装专用软件；而CIFS集成在OS内部，无需额外添加软件；

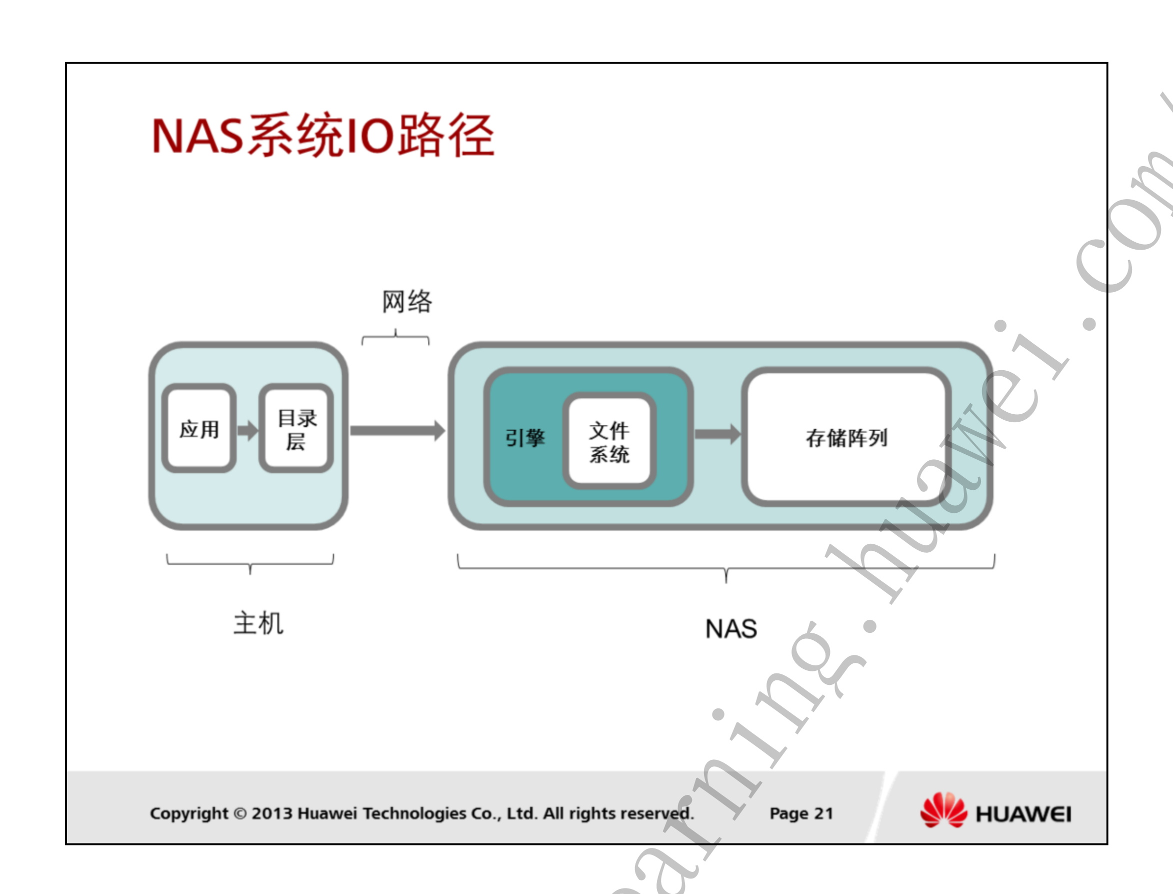
3、NFS属无状态协议，而CIFS属有状态协议；NFS受故障影响小，可以自恢复交互过程，CIFS不行；从传输效率上看，CIFS优于NFS，没有太多冗余信息传送；

4、两种协议都需要文件格式转换，NFS保留了unix的文件格式特性，如所有人、组等等；CIFS则完全安装win的风格来作。

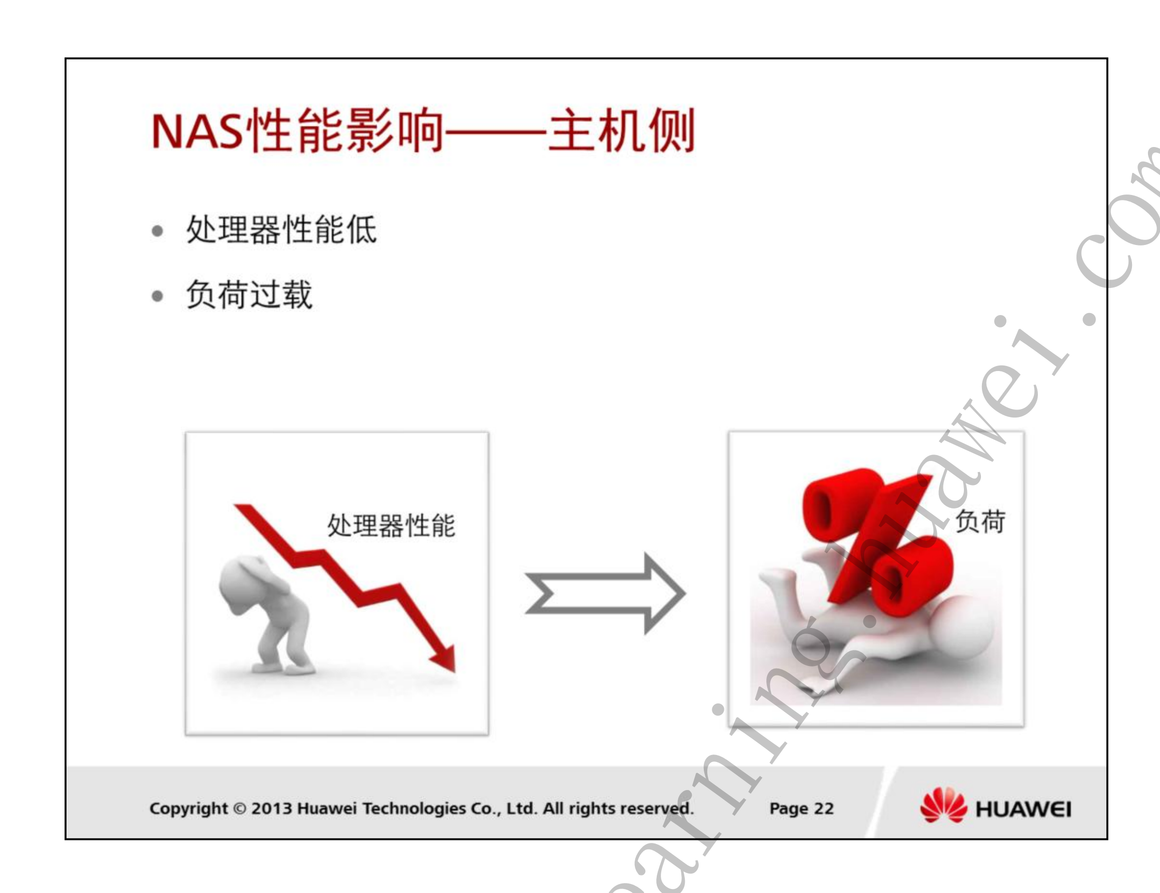


目录

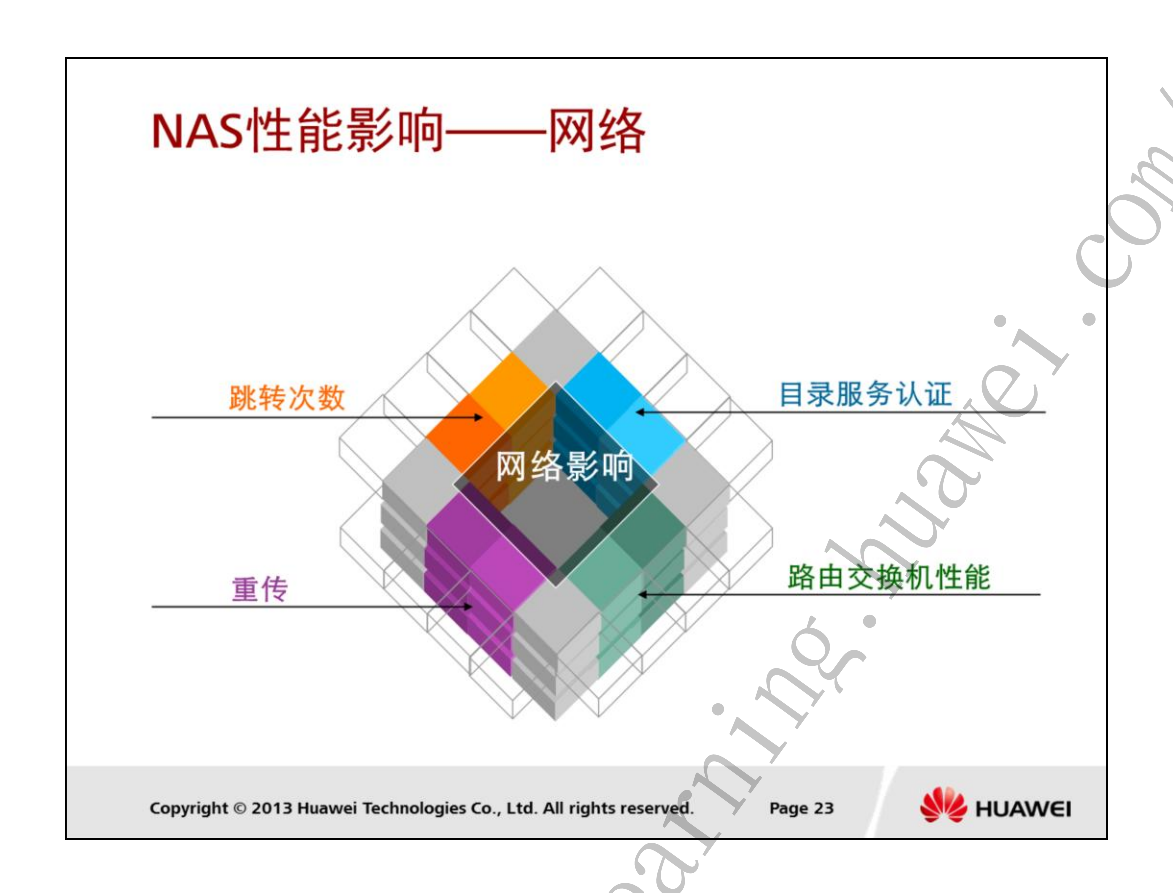




在NAS系统中，NAS引擎通过网络将后端的存储资源以文件夹的形式对外提供，当客户端在访问使用NAS存储资源的时候，一般由应用发起I/O操作，然后通过网络到达NAS引擎，引擎再对I/O操作进行处理，最后命令到过存储阵列，完成IO操作。



访问主机客户端本身配置较低，操作系统要花较多的时间来处理应用，如果再加上主机上运行的业务程序过多，那么需要更多的时间去处理接收到的响应，那么就会影响到主机端应用对存储的操作。



- 跳转次数

  大量的网络包跳转会增加延迟；

- 重传

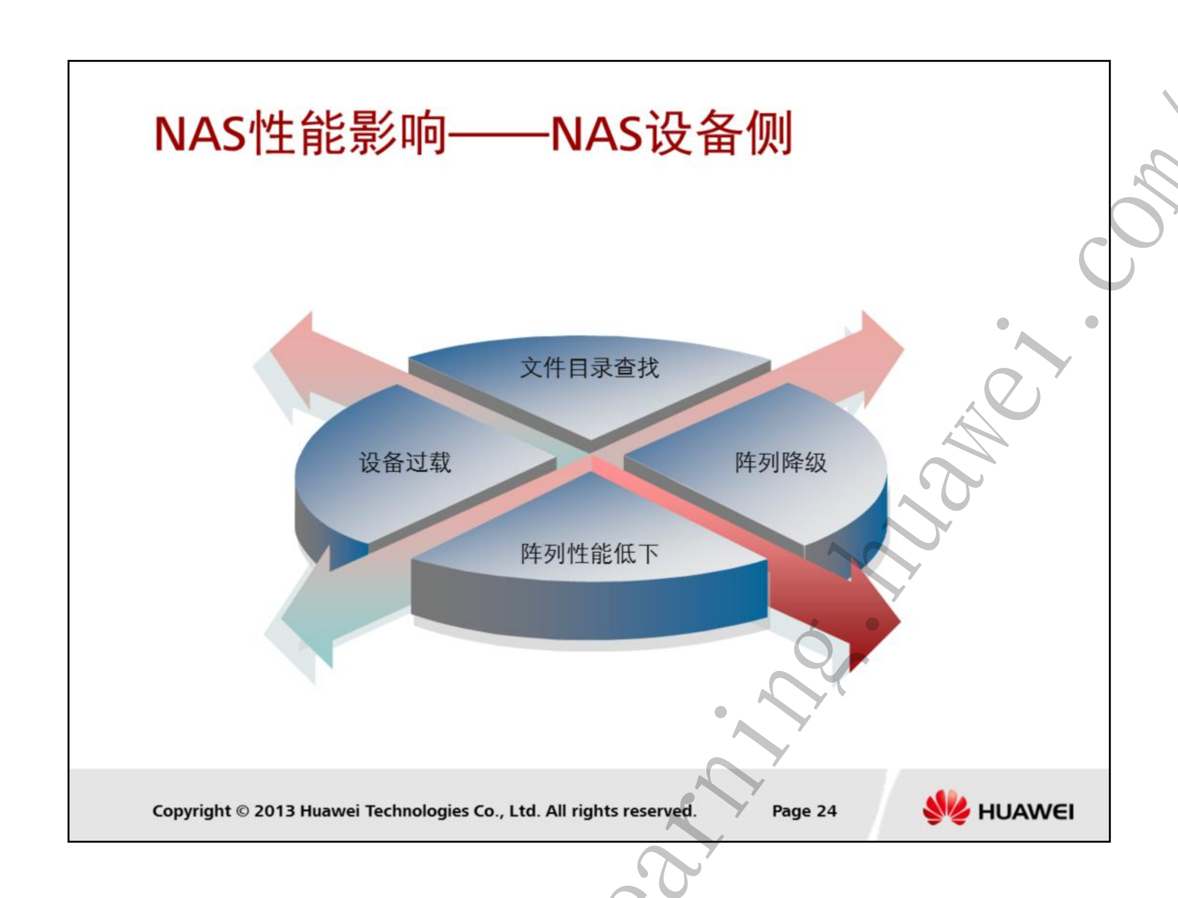
  链路错误、缓冲区溢出和流量控制机制都会导致重传。这会导致未到达指定目的地的报文被重新发送。在配置网络设备的速率参数、双工通信参数以及NAS头参数时要注意使它们彼此匹配。不恰当的配置会导致错误和重传，增加延迟。

- 目录服务的认证

  例如LDAP、活动目录或NIS：认证服务是网络上必须的服务，而且必须拥有充足的带宽和足够的资源来支持认证过程产生的负载。否则，大量的认证请求发向服务器会增加延迟。当然，只有当认证发生时才会增加延迟。

- 路由交换机性能

  在网络中，一个过载的设备所需要的响应时间总是比优化状态下使用的或低负载使用的设备所需要的响应时间要长。



- 文件目录查找

  文件目录过大过深，在查找的时候会非常耗资源，对NAS性能造成一定影响；

- 设备过载

  设备过载，会长时间处于高负荷的状态，对应用的响应造成影响；

- 存储阵列降级

  阵列降级状态，阵列内的磁盘参与重构，一般重构数据量都比较大，直接影响阵列的性能，进而影响NAS的性能；

- 存储阵列设备性能低下

  存储阵列设备性能低下会直接影响NAS的性能，因为NAS对数据的IO操作最终的处理设备还是在阵列上进行的，如果阵列性能低下了，整个NAS性能性能也就低下了。



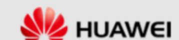
目录



IO类型

- CPU密集
  - 应用程序内部逻辑复杂，占CPU资源多，对磁盘的实际IO操作较少。
- IO密集
  - 应用程序内部逻辑简单，占CPU资源不多，但存储数据量比较大并且较频繁。

Page 26



要对SAN和NAS进行性能比较，首先要分场景，它们有各自所擅长的场景，不同的场景所对应的就是IO的类型，从大的方面分，可把IO类型分为CPU密集型和IO密集型，当然，还有其它的一些IO特征，比如IO的数据大小，是随机的还是连续的，读写所占的比例等，这些都会对IO的性能造成一定的影响。

SAN与NAS比较

比较项	速度	成本	客户端资源占用	协议	共享	适用场景
SAN	快	高	无文件系统，占用客户端资源提供文件系统功能	FC,SCSI, iSCSI	需客户端安装专用共享文件系统，不易共享	大块连续IO密集环境
NAS	慢	低	有自自己的文件系统少，客户端资源占用少	CIFS,NFS, HTTP, FTP	自有文件系统能直接提供共享	小块CPU密集环境

 Page 27



SAN存储提供数据块级别的访问，需要客户端对使用的存储资源进行格式化，然后才能使用，由于SAN网络一般采用FC网络架构，是专用网络，速度较快，适用于IO操作较密集的环境；

NAS提供的是文件级别的共享，用户能像使用网上邻居共享的文件夹一样使用，使用比较方便。而且通过TCP/IP网络共享，在理论上共享的距离不受限制，适用小块CPU密集环境。



目录





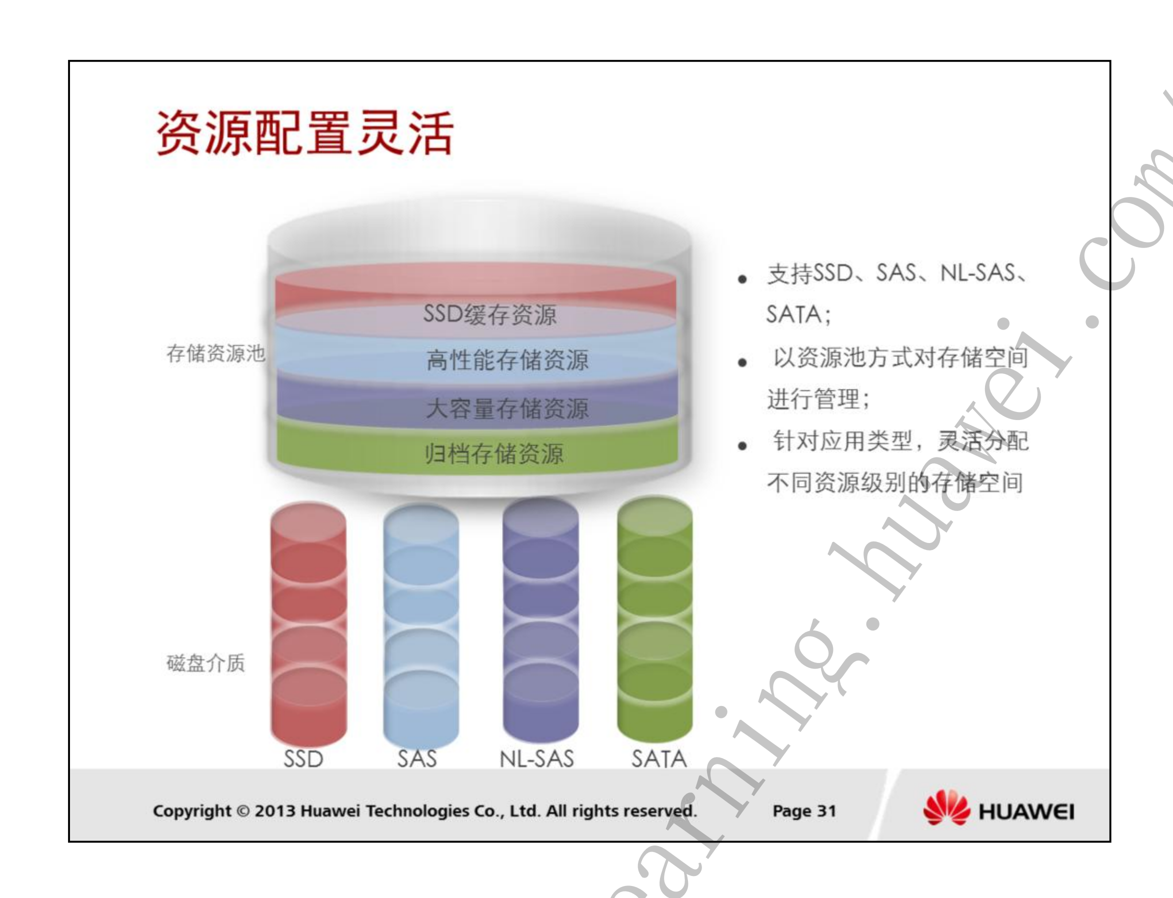
N8500是一款集群化的中高端NAS存储系统，针对高效数据共享产品的需求，具有灵活的横向及纵向可扩展性；可用于金融、政府、石油天然气、健康和生命科学、制造业、E-Discovery等行业。

N8500的特点主要体现在高可靠性、高性能、支持动态分级存储、支持业务网口绑定、支持用户配额管理、支持文件系统快照、支持基于LAN的备份、支持文件系统多重镜像、支持文件系统在线扩容、支持NFS和CIFS协议的共享、支持FTP协议访问、支持域环境、支持存储单元后台格式化。



我们先看一下传统NAS的架构，传统NAS非集群架构，多为Active-Standby架构，NAS引擎各自管理被分配的空间，如左图，文件系统A仅属于A控管，文件系统B仅属于B控管，AB控相互独立。前端服务器需要访问文件系统A，只能从A控获取，可以想象，随着服务器增多，A控制器的压力逐渐增大，形成瓶颈。当A控制器故障的时候，控制器B才会接管文件系统A。

N8500采用集群架构，所有引擎均处于活动状态，NAS引擎管理后端整个存储空间，文件系统A的数据，服务器群既可以通过A控访问，也能通过B控访问，文件系统B也是如此。通过多个控制器访问同一文件系统，提升并发处理效率，相当于N个人同时干一个活，在HPC，媒资等高性能存储市场效率可成倍提升。同时支持故障自动切换，保障业务持续运行。



对于存储空间的管理，N8500通过存储池方式对不同类型磁盘介质进行统一管理，可支持SSD、SAS、NL-SAS、SATA多种磁盘类型，可针对不同类型的应用，划分不同级别的资源空间，这样做可以实现按需分配，灵活调度，达到资源的最优配置。

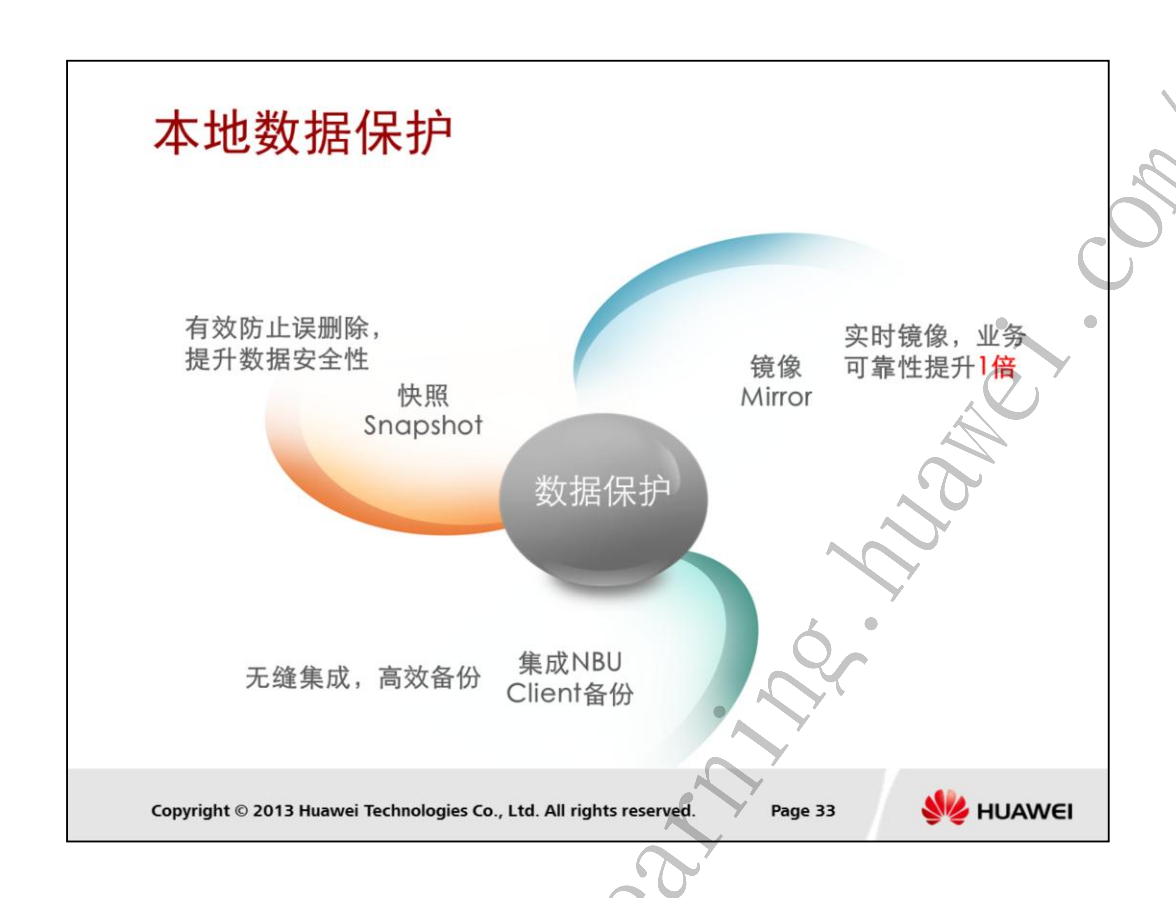


智能迁移，提升性能并降低采购成本。大家都知道，数据可以根据其被使用的频度分为热点数据和非热点数据，N8500通过SmartCache技术和DST技术实现对热点数据的加速，对非热点数据的智能迁移。

DST预测功能是通过对用户当前业务负载特点的分析，将负载高的数据预测分配到高性能层或者性能层，将负载非常低或者完全为0的数据预测分配到容量层，在满足用户当前业务性能需求且保证总价格最低的情况下，给出用户存储池中的所有业务数据在各层分布最合理的预测比例，以指导用户后续进行合理的配置。

SmartCache技术通过对热点数据进行智能迁移，将热点数据镜像至SSD缓存池中，缓存池中始终保留最热点的数据，提升热点数据的访问性能，通常可提升5倍或以上。广泛适用于互联网，运营商彩铃彩信等应用场景。

而对于非热点数据的智能迁移，N8500通过DST技术来实现，通常热点数据和非热点数据大约是2:8，将大量的非热点数据根据策略自动迁移至廉价的大容量SATA盘上，降低采购成本，通常可降低60%的硬盘采购成本。相比业界的分级存储，DST技术的迁移对象是针对文件进行迁移，热点识别更精准，迁移粒度更灵活。广泛适用于音视频点播，邮件系统，媒资库等存在大量“冷”数据的应用场景中。

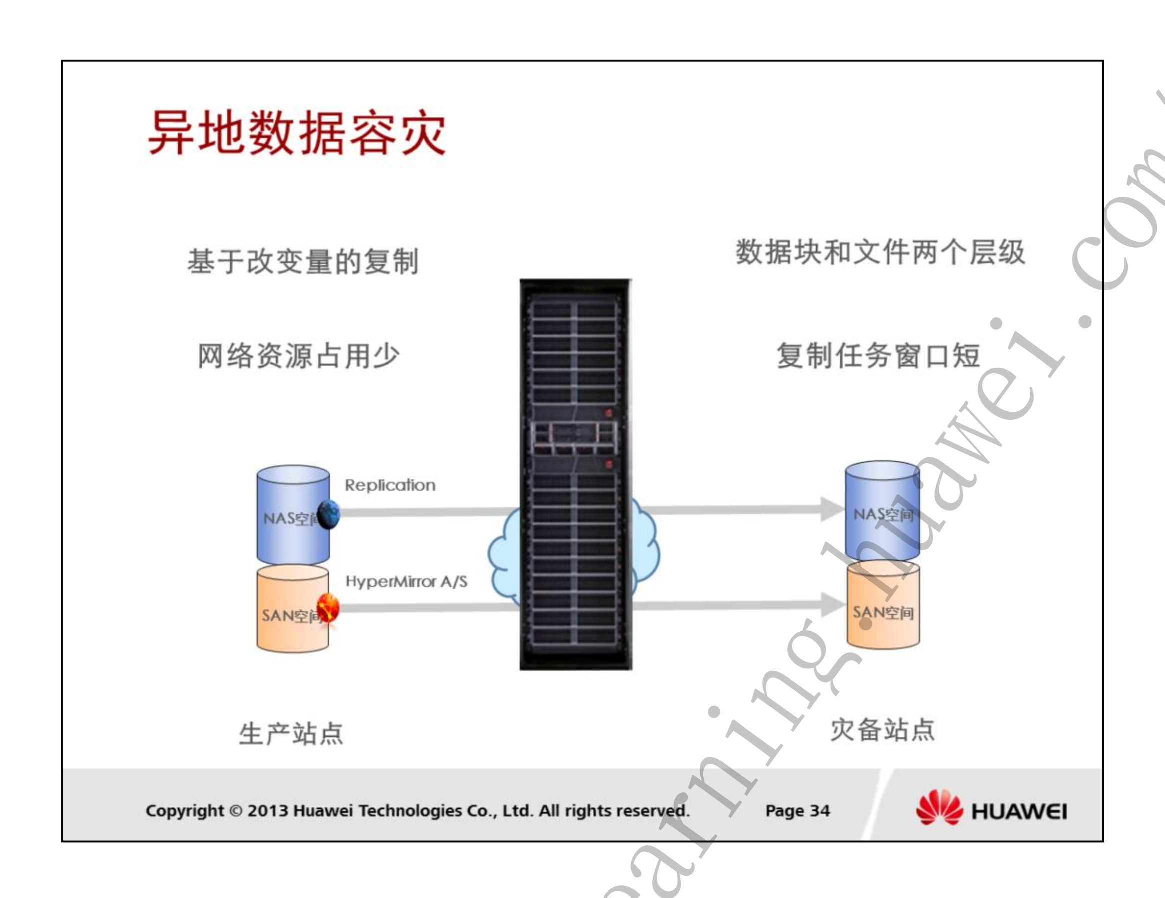


前面我们解决了存储的性能，扩展性，空间利用率，那么如果保证数据的安全呢，N8500通过多种技术保证数据安全。

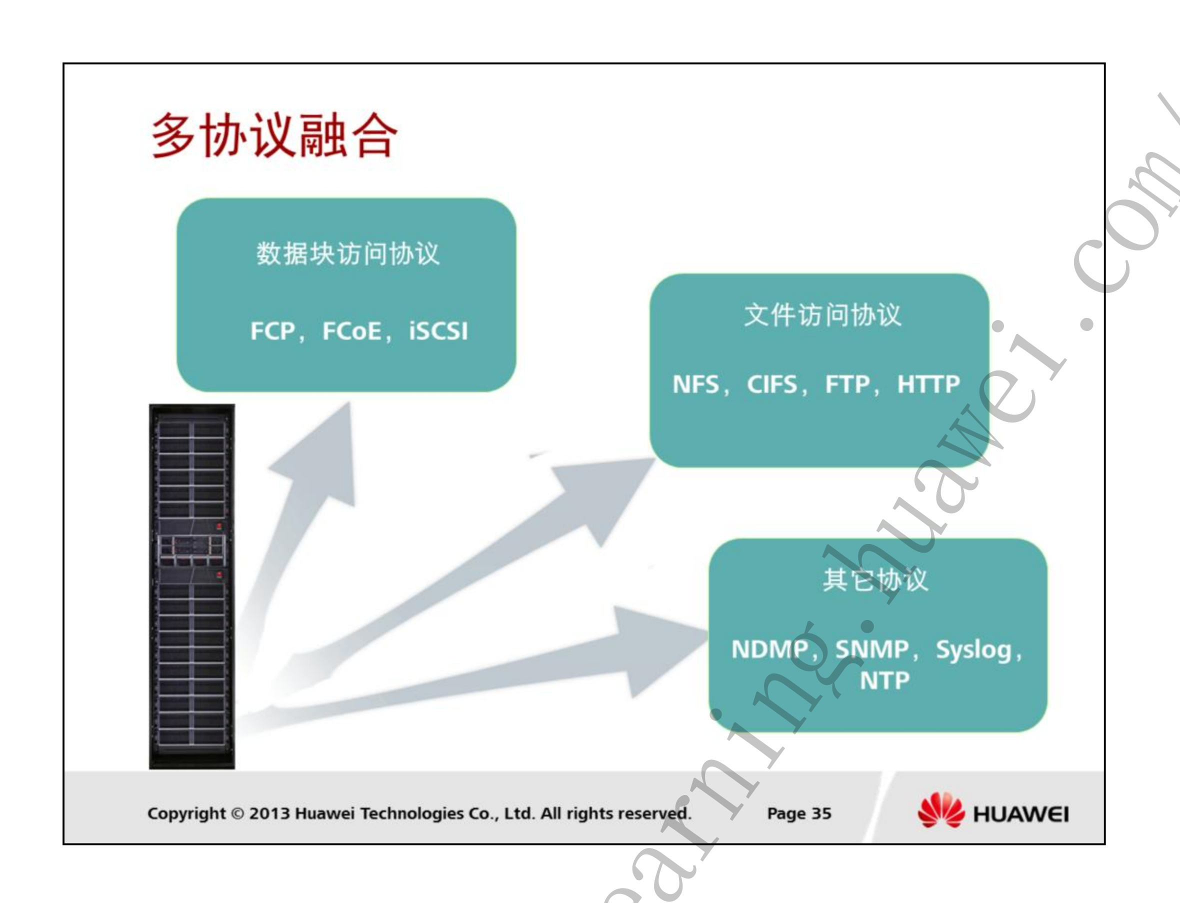
快照技术能有效解决数据的误删除，提升数据安全性。

镜像技术，N8500后端连接多套存储单元，数据同步镜像写入存储单元，一个存储单元故障时不影响业务运行，提升业务可靠性。

集成NBU的备份客户端，实现数据高效备份,NBU是赛门特克公司备份软件Netbackup的简称。

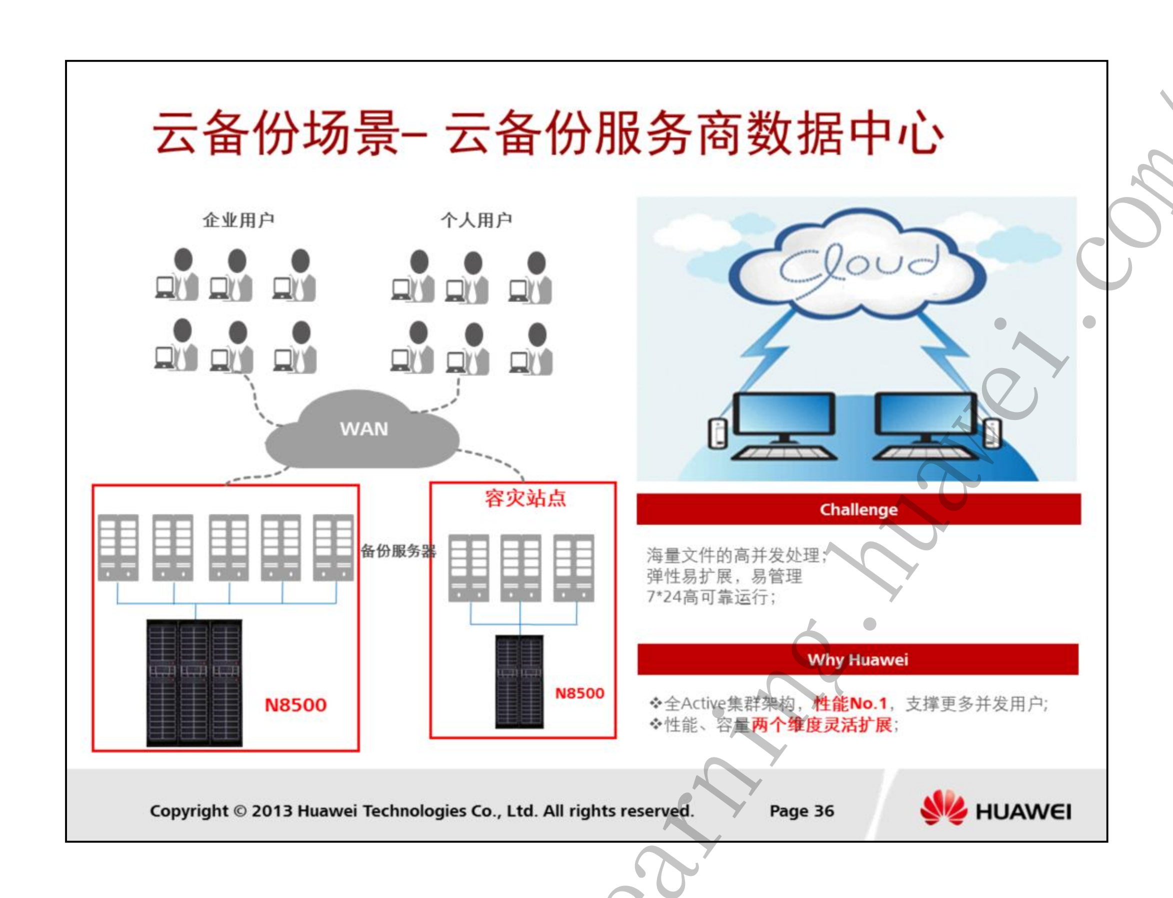


有了本地的数据安全保护还不够，对于关键数据，有时我们还需要在异地保留一份数据。N8000支持基于SAN空间和NAS空间的两个层级数据复制，进一步提升数据的安全性。远程复制是基于改变量的复制，占用网络资源少，缩短复制任务窗口。生产站点数据一旦发生丢失，可通过灾备站点对数据进行恢复。

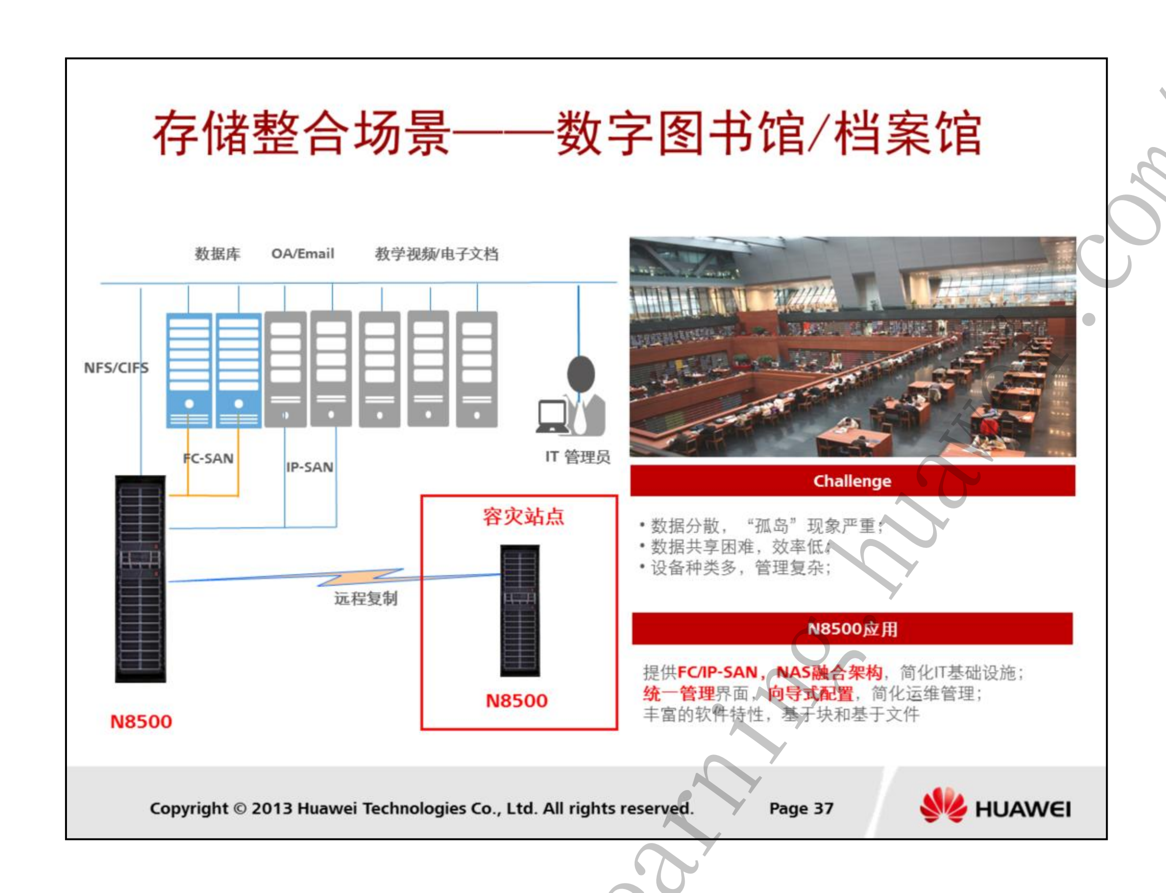


多协议融合，N8500支持FCP，FCoE，iSCSI协议，相比在文件系统层面上再虚拟SAN空间的方式，有更高的效率；支持基于文件的NFS，CIFS，FTP，HTTP访问协议，同时支持NDMP，SNMP，Syslog等其他协议。通过将不同协议融合至一套存储系统中，满足不同应用类型对存储的要求，简化IT基础架构，减少设备数量和采购成本。

NDMP是由Network Appliance Corporation和Legato System公司合作开发的一种基于NAS的容灾备份技术。它提供了一个开放的协议，利用NAS设备进行数据的备份/恢复，符合NDMP的备份应用程序可以通过它来控制任何运行NDMP服务器应用程序的NDMP主机的备份与恢复。



云备份应用场景主要关注性能，扩展性和对海量文件的处理能力，华为N8500采用全Active集群架构，适合海量用户并发访问存储，同时随着用户数量增大，数据容量不断增大，N8500也能轻松扩容应对，能有效保护原有投资。



存储整合应用场景主要关注存储能提供丰富的接口和协议，提供丰富的软件特性，满足不同应用程序，不同服务等级对存储的不同需求，在数据中心放置一套存储就能满足各种需求。华为N8500提供多种接口模式，多种协议类型对外提供存储资源，能很好的满足存储整合应用场景的需求。



总结

- NAS组成结构及实现
- NAS文件共享协议NFS以及CIFS
- NAS系统文件IO与性能
- SAN与NAS的区别及联系
- 华为NAS产品



思考题

1. 什么是NAS
2. NAS的硬件组成及软件架构
3. NFS,CIFS的应用
4. SAN与NAS的比较
5. 华为的NAS产品



练习题

- 多选题

  1、NAS的组件由哪几部分构成？（　　　）

  A、存储

  B、网络

  C、引擎

  D、服务器

  2、以下哪些选项会影响NAS设备性能（）

  A、文件目录过大过深

  B、设备过载



练习题

C、存储阵列降级

D、存储阵列设备性能低下

3、NAS较文件服务器的优势有（ ）

A、稳定性、可靠性高

B、数据安全性高

C、存储空间大

D、具有容灾、备份设计

 Page 41



答案

1、ABC

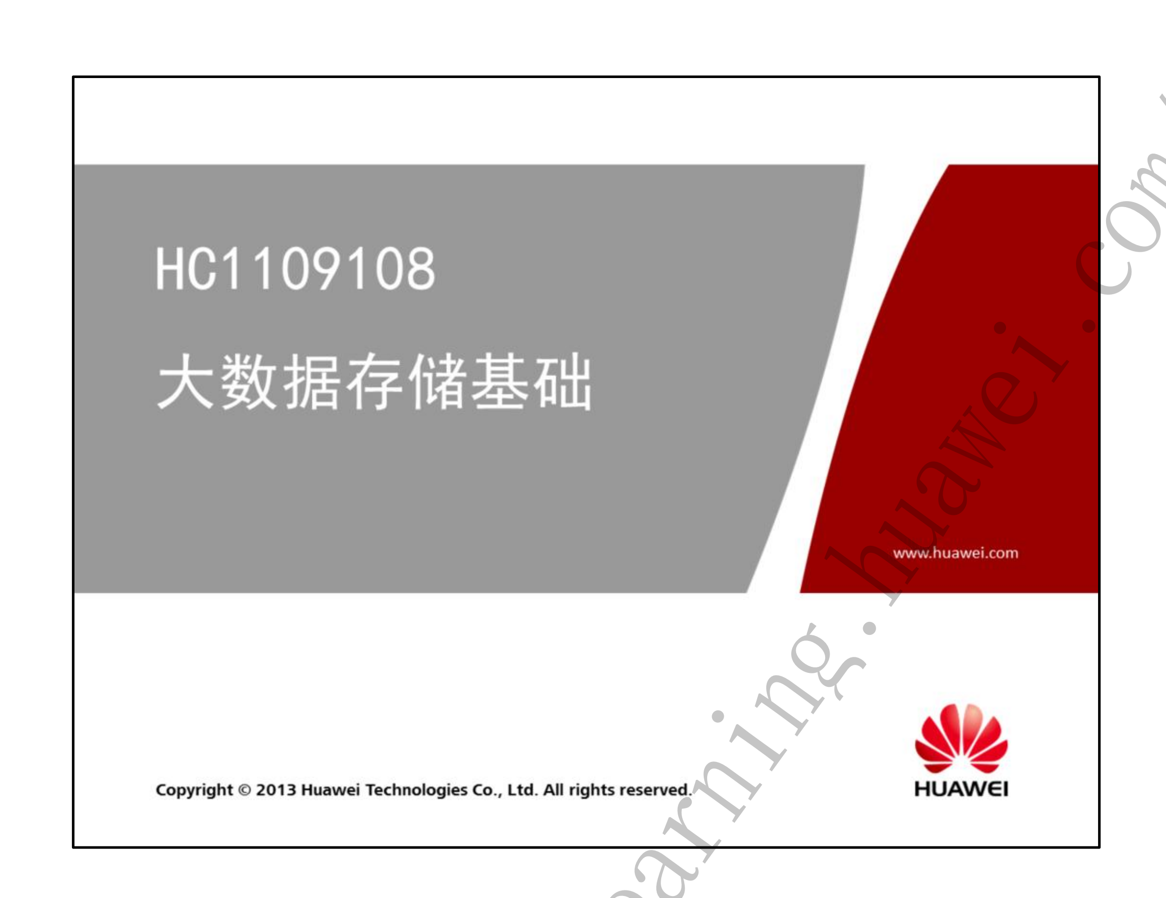
2、ABCD

3、ABCD


Thank you
www.huawei.com
Copyright © 2013 Huawei Technologies Co., Ltd. All rights reserved.
Page 42
HUAWEI

HC1109108 大数据存储基础




HC1109108
大数据存储基础
www.huawei.com
Copyright © 2013 Huawei Technologies Co., Ltd. All rights reserved.
HUAWEI

目标

- 学完本课程，您将能够：
  - 了解大数据的基本概念
  - 理解对象存储技术
  - 了解大数据的处理技术
  - 掌握华为大数据解决方案及技术





目录

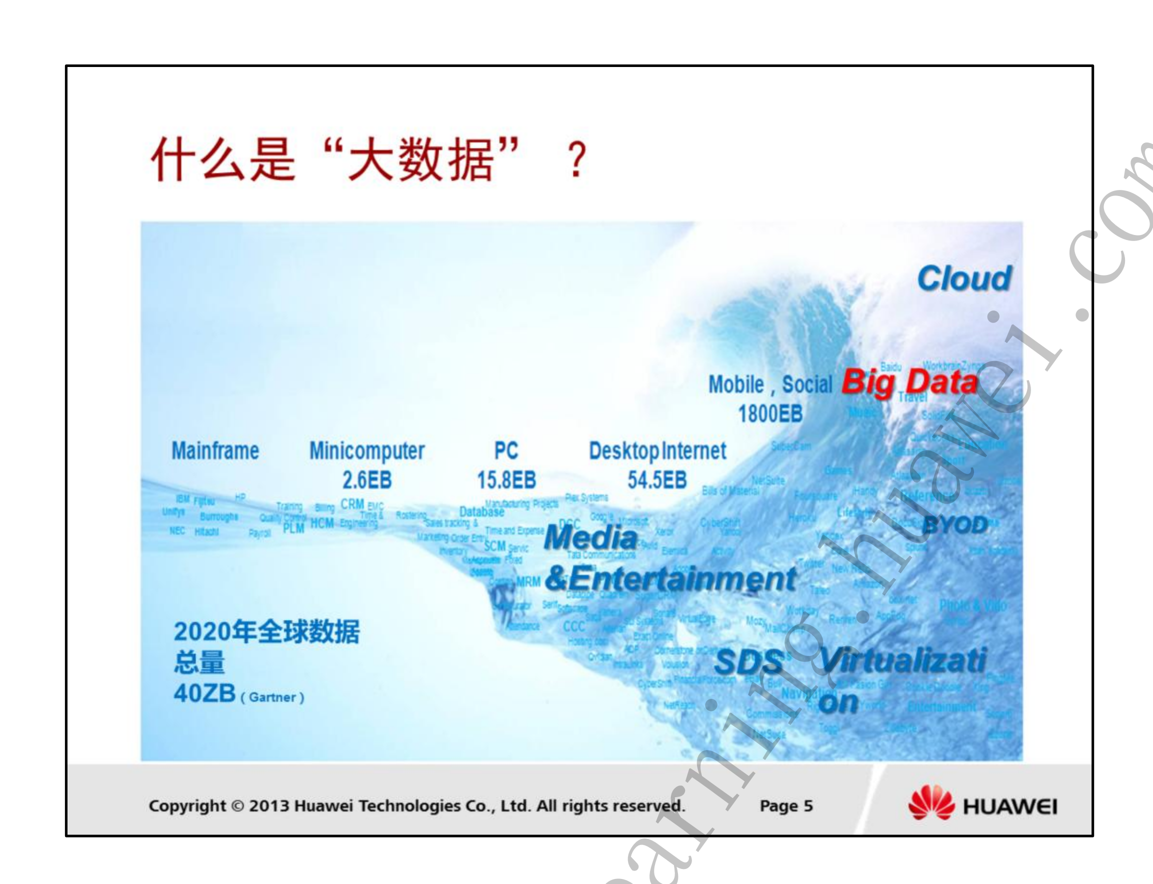




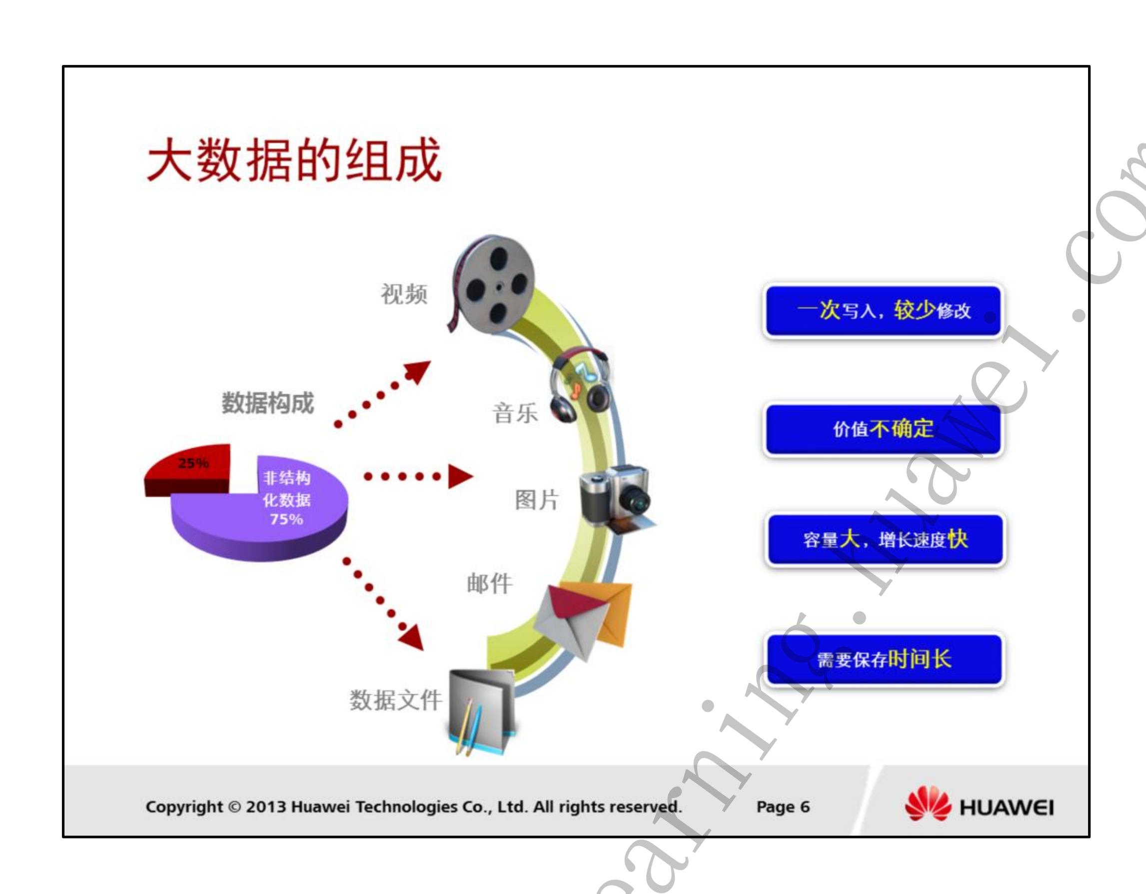
从世界诞生之日至2003，人们共创造出5EB，而2011年只需要不到2天便能产生如此多的信息量，而2013年将只需要10分钟，而到2020年，每年创造的数据容量将高达35ZB，导致数据爆炸的原因是多方面的：

- 智能终端普及，如手机，Ipad；
- 人们分享与沟通需求增加，如微博，微信，Facebook等应用的流行；
- 网络高速发展；
- 数据结构发生变化，视频、图片占绝大部分；
- 高清视频、图片的大量产生；

以上诸多原因，导致海量数据的存储与管理成为所有IT系统必须面对的问题。



所谓“大数据”，就是用现有的常规软件（企业中的关系型数据库），在可容忍的时间内无法处理的数据的集合。



研究表明，数据的75%以上是非结构化数据，主要由视频、音乐、图片、邮件、数据文件构成。

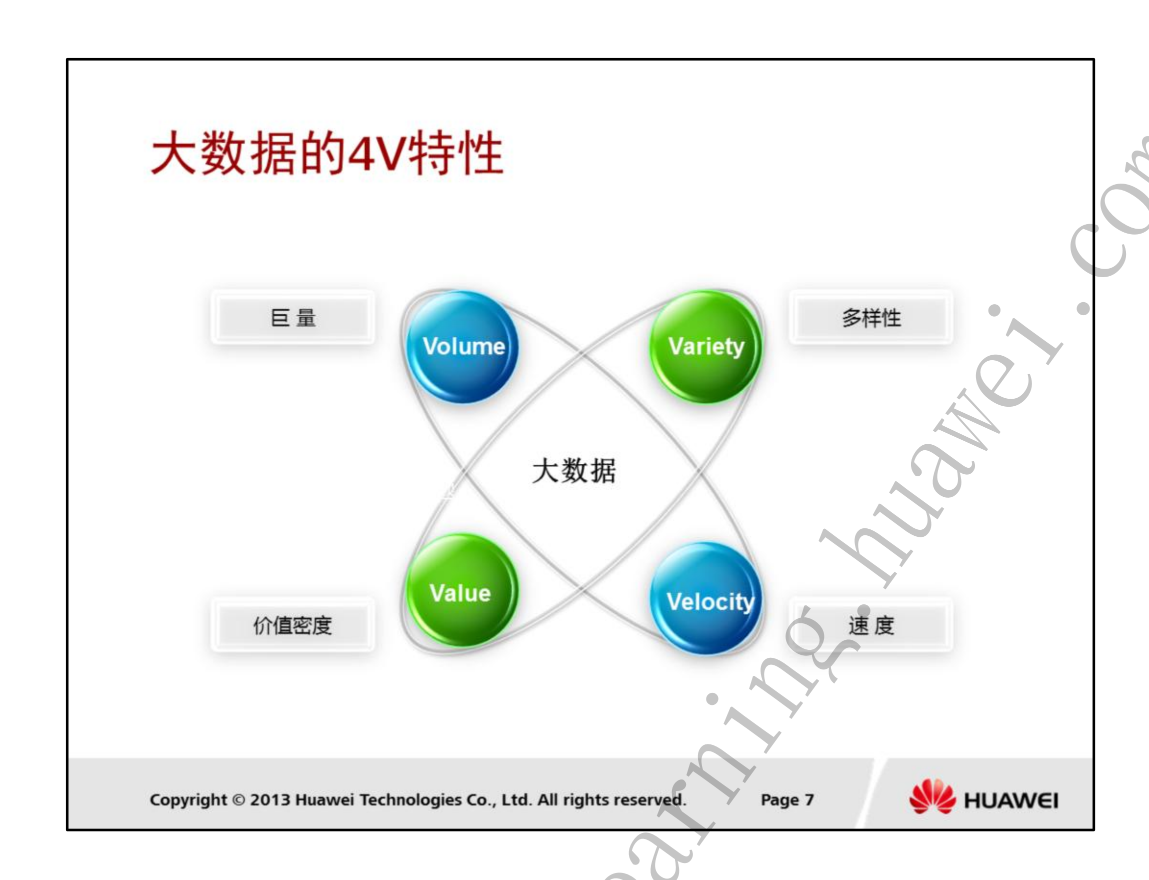
绝大多数的海量数据具有如下特征：

1、一次写入，较少修改。如大量的视频、图片文件，保存之后主要是读取，很少编辑；

2、数据价值不确定。如照片和视频，可能因为某个偶然事件，而使价值增加。如某人在某一段时间内成为焦点人物，那么他小时候的照片就变得有价值。视频监控数据也有类似特征。你不知道这些数据什么时候有用，但你不能轻易把它扔掉；

3、容量大，增长速度快；

4、需要保存时间长；有些数据可能要保存几十年，甚至更长时间。



- Volume:全球在2010年正式进入ZB 时代，IDC预计到2020 年，全球将总共拥有40ZB 的数据量。
- Variety:如今的数据类型早已不是单一的文本形式，订单、日志、音频等结构化数据、半结构化数据和非结构化数据共存；非结构化数据快速增长，总量占到总数据量80%-90%；
- Value:沙里淘金，价值密度低，一部数小时的视频，有价值的数据可能仅仅只有一两秒；
- Velocity:数据产生和更新频率极高，非结构化数据产生速率比结构化数据快10-50倍，是传统数据仓库的10-50倍；Facebook每天产生50TB的日志数据。

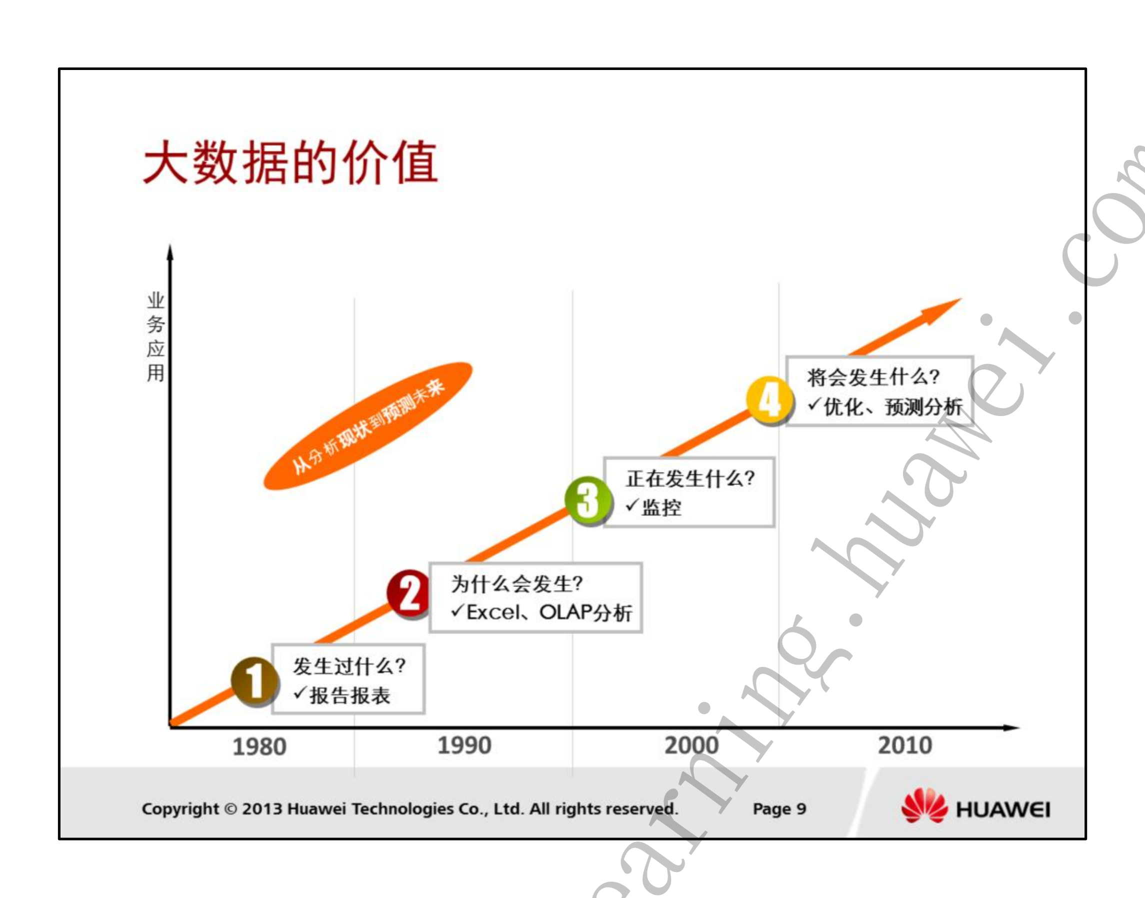


- 数据多源化

如今年的数据已不象以前一样只产生于某些领域，安装在汽车上的GPS系统，人们日常使用的手机等，都是数据产生源头。

- 软硬件技术发展
  - 计算机的计算速度按照摩尔定律逐年增加，存储介质的存储密度大幅提升，单位存储容量的成本逐年降低；
  - 大量的非结构化数据是大数据区别于传统数据的一个主要特性；传统的关系形数据库无法处理急剧增加的非结构化数据，基于Hadoop的分布式数据处理技术不断浮现， 使非结构化数据处理成为可能。
- 云计算普及

大数据处理需要性能强大的软硬件处理平台，搭建这样的平台是一笔不小的开支，对于一些规模较小企业可能无法承担。云计算的Iaas和PaaS服务可以为企业提供按需付费的软硬件平台服务，通过这种方式，小企业同样可以实现大数据处理。



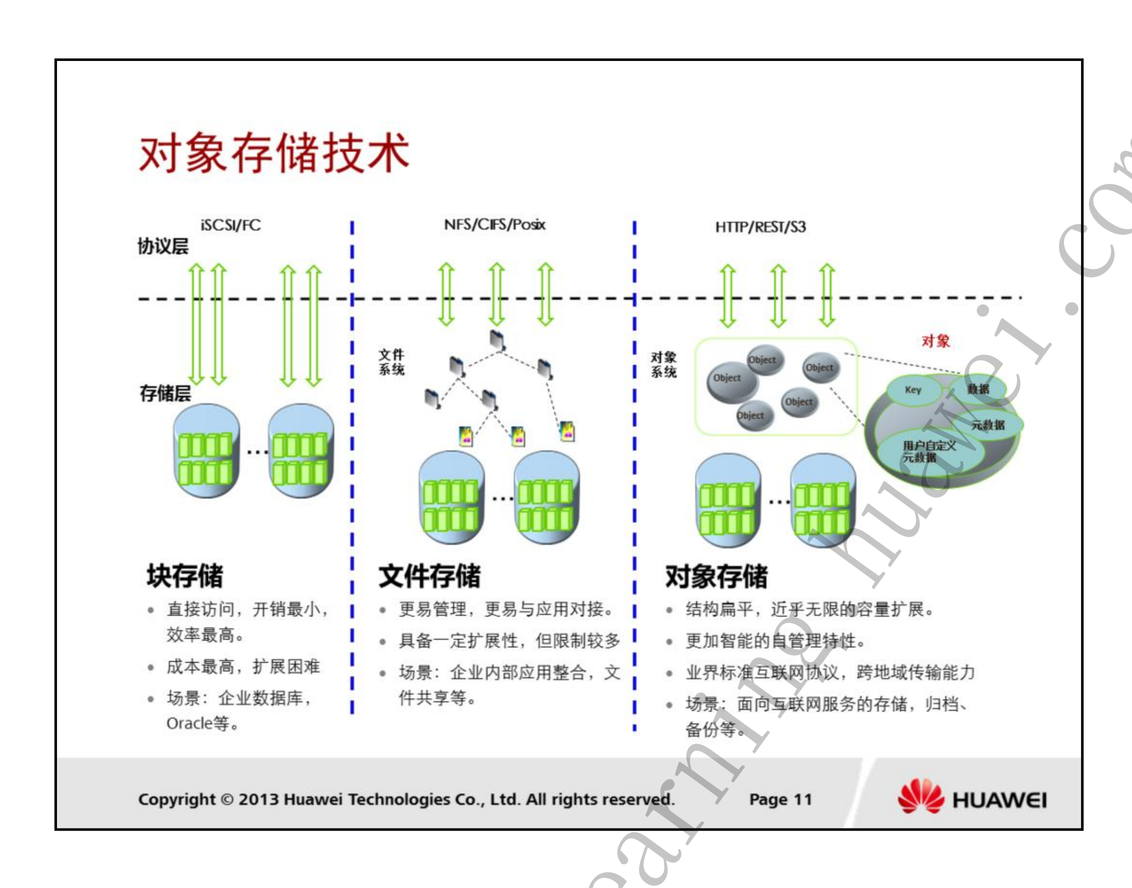
大数据以前，人们对数据的利用主要是分析过去到现在发生了什么、为什么会发生，并做出报告。大数据的核心在于“预测”，预测将基于 “数据之间的关联性” 而非“为什么是这样的因果性”，我们只需要按照预测出来的趋势去响应，使用这些结果。

通过对大量数据的分析，从中发现数据中蕴含的有价值的规律，从而预测未来，是一种非常有意义的活动。例如：Google 早在 2008 年推出了“流感趋势”网站。它建立的假设基础是：人们在遭受疾病困扰时，会比在身体健康时，花更多的时间搜索疾病相关内容。因此，通过分析一个国家，在特定时期的流感相关搜索量，便可以推算出病毒的传播情况。这个预测非常靠谱，通常与美国疾病控制和预防中心（CDC）的数据相差无几。事实上，有几次传染病初期的预测，甚至比 CDC 早了一周之久。众所周知，疾病初期预测将为政府及时采取部署，提供了有利的时机。



目录



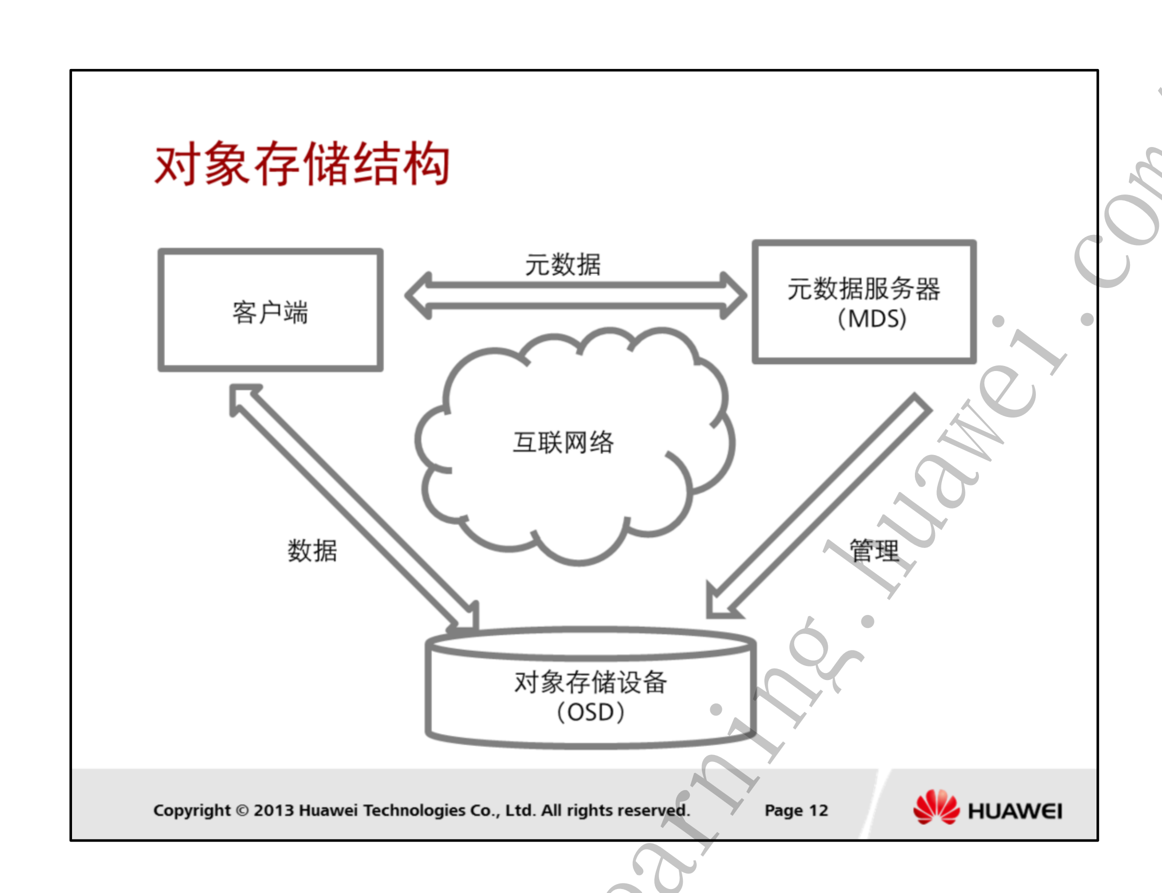


对象存储是一种新兴存储技术，对象存储系统综合了NAS和SAN的优点，同时具有SAN的高速直接访问和NAS的数据共享等优势，提供了高可靠性、跨平台性以及安全的数据共享的存储体系结构。对象存储与块存储、文件存储的对比如下：

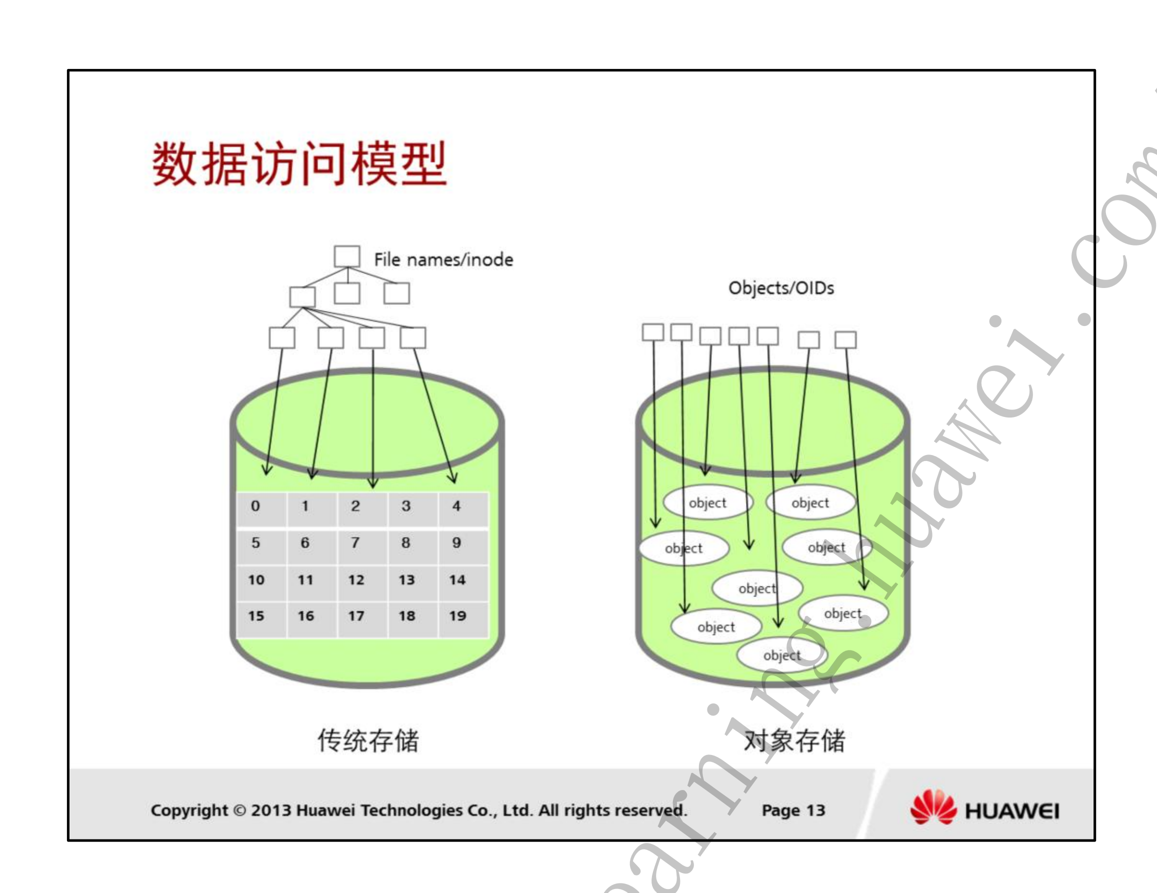
1、块存储对存储层直接访问，开销最小，效率最高，速度最快。但成本最高，扩展困难。块存储采用iSCSI/FC协议，很难跨网络传输。适合的应用场景是企业数据库，如运行Oracle等；

2、文件存储是在块存储之上构建了文件系统，采用目录-目录-文件的方式组织数据，更容易管理。因为大多数应用程序都是对文件进行操作，因此文件存储更容易和应用系统对接。文件系统受目录树的限制，扩展性受限，一般最多扩展到几十PB。文件系统适用于企业内部应用整合，文件共享场景；

3、对象存储是在块存储之上构建了对象管理层，与文件系统相比，对象系统层是扁平的，扩展限制少，因此拥有近乎无限的扩展性。对象由唯一的Key，文件，数据（文件），元数据，自定义元数据构成，由于包含了自管理信息，更加智能。对象存储采用兼容标准的互联网协议接口，可以跨地域传输。对象存储适用于面向互联网服务的存储场景，以及企业内部的归档、备份场景。



在对象存储系统中，元数据服务器（MDS）承担起了文件与对象存储设备（OSD）的映射，文件与目录的组织关系任务。它提供了所有文件系统名字域操作，例如文件查找，文件创建，文件和目录属性处理。从客户端角度去看元数据服务器好似是文件的逻辑窗口，而OSD就是文件的物理窗口。当用户对某个文件进行操作，首先文件系统从MDS上获取到文件的实际存储地址，然后根据这个地址到OSD上进行存取操作，后续的I/O操作都不需要再访问MDS，这样减少了MDS的负担，从而为系统的扩展提供了可能性。



传统存储文件系统采用树目录的形式，当文件层数较多，且文件数量比较大的时候，对根节点的压力比较大，要去查找一个文件耗时较长，影响性能。对象存储采用扁平化的结构，采用去中心化的思想，再多的文件在访问的时候几乎不受影响，系统能很方便的扩容。

对象存储的优势

- 采用对象接口，灵活分割数据
- 对象扁平化，易访问扩展
- 自动化管理
- 多租户
- 数据完整性和安全性

 Page 14



- 采用对象接口，灵活分割数据

  对象存储对外提供更抽象的对象接口，而不是SCSI或文件接口。与SAN存储以逻辑扇区为单位的较细粒度的固定IO（512B~4KB）不同，对象存储IO粒度更有弹性，支持几个字节(B)到数万亿字节(TB)范围内的任意对象大小，使得业务可根据需要灵活的分割数据。

- 对象扁平化，易访问扩展

  扁平化的数据结构允许对象存储容量从TB级扩展到EB级，对象存储系统通常在一个横向扩展（或网格硬件）架构上构建一个全局的命名空间，这使得对象存储非常适用在云计算环境中使用。某些对象存储系统还可支持升级、扩容过程中业务零中断。

- 自动化管理

  对象存储支持从应用角度基于业务需求设置对象/容器的属性（元数据）策略，这使得对象存储具备云的自服务特征同时，有效的降低运维管理的成本。

- 多租户

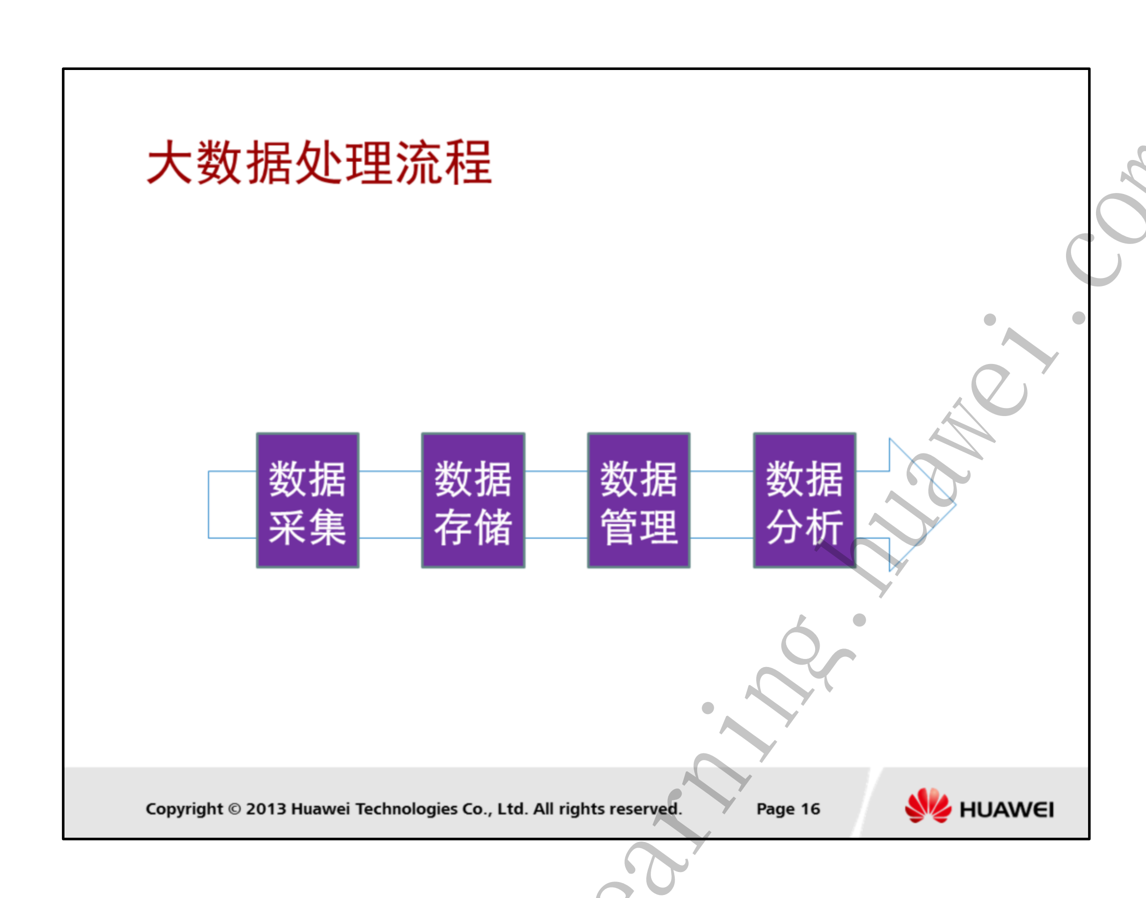
  多租户特性可以使用同一种架构,同一套系统为不同用户和应用提供存储服务,并分别为这些用户和应用设置数据保护、数据存储策略，并确保这些数据之间相互隔离。

- 数据完整性和安全性

  对象存储系统一般通过连续后台数据扫描、数据完整性校验、自动化对象修复等技术，新型的技术应用大大提高数据的完整性和安全性。

华为N9000设备采用的就是对象存储的技术，提供大数据的解决方案，在后面深入的介绍。


目录
1. 大数据的基本概念
2. 对象存储技术
3. 大数据处理技术
4. 华为大数据实践
Copyright © 2013 Huawei Technologies Co., Ltd. All rights reserved.
Page 15
HUAWEI



针对大数据处理的流程与一般数据处理流程基本一致，包括数据采集、数据存储、数据管理、数据分析四个环节。

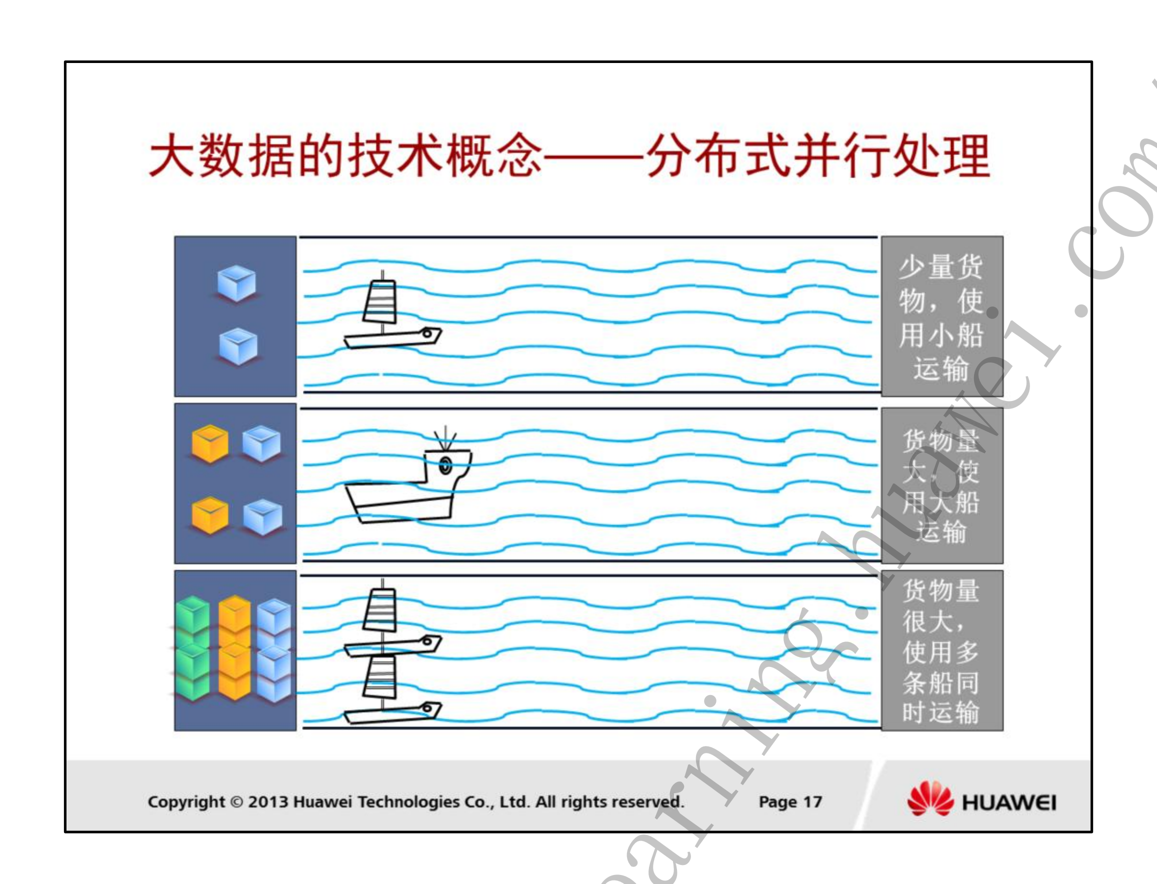
数据采集是利用多种途径、方法和工具来获取所需要的数据，为后续数据分析提供依据。在大数据的背景下，需要采集什么数据？通过什么途径采集数据？运用什么方法采集数据？利用什么工具实现更高效的数据采集？等都是我们必须关注的问题。

数据存储是对采集到的数据在进行数据分析前实现数据的传输和存储。随着大数据背景下数据量级数级增长，传统的数据存储方式已经很难承载大数据存储的需求，因此需要采用更新的技术来实现数据存储。

数据管理是数据存储的延伸。在数据存储的基础上对数据进行深加工，进一步实现数据细分，为后续数据分析提供直接可用的元数据，提升数据分析的效率。

数据分析是利用数据分析的方法、模型、工具对数据进行分析，最终得出分析结果，以此满足大多数常见的分析需求。基于前面分析和查询的数据进行更深入的数据挖掘，可以满足更高级别的数据分析需求。

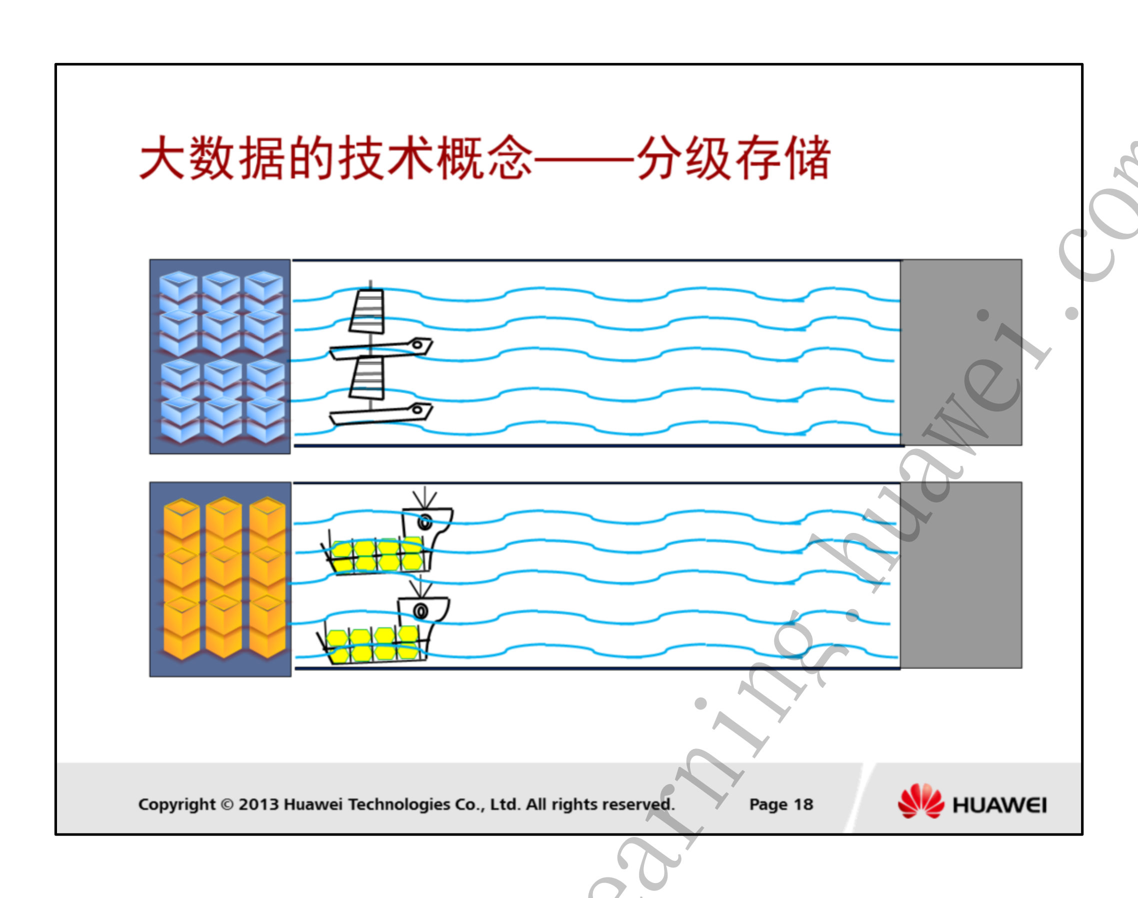
本章重点围绕数据存储以及数据存储的延伸——数据管理来介绍大数据的关键技术。



在介绍大数据关键技术前，为便于理解，我们先通过简单的图表形式来介绍大数据的一些技术概念。这里我们通过河流的船运来描述从“数据”到“大数据”的技术变化。

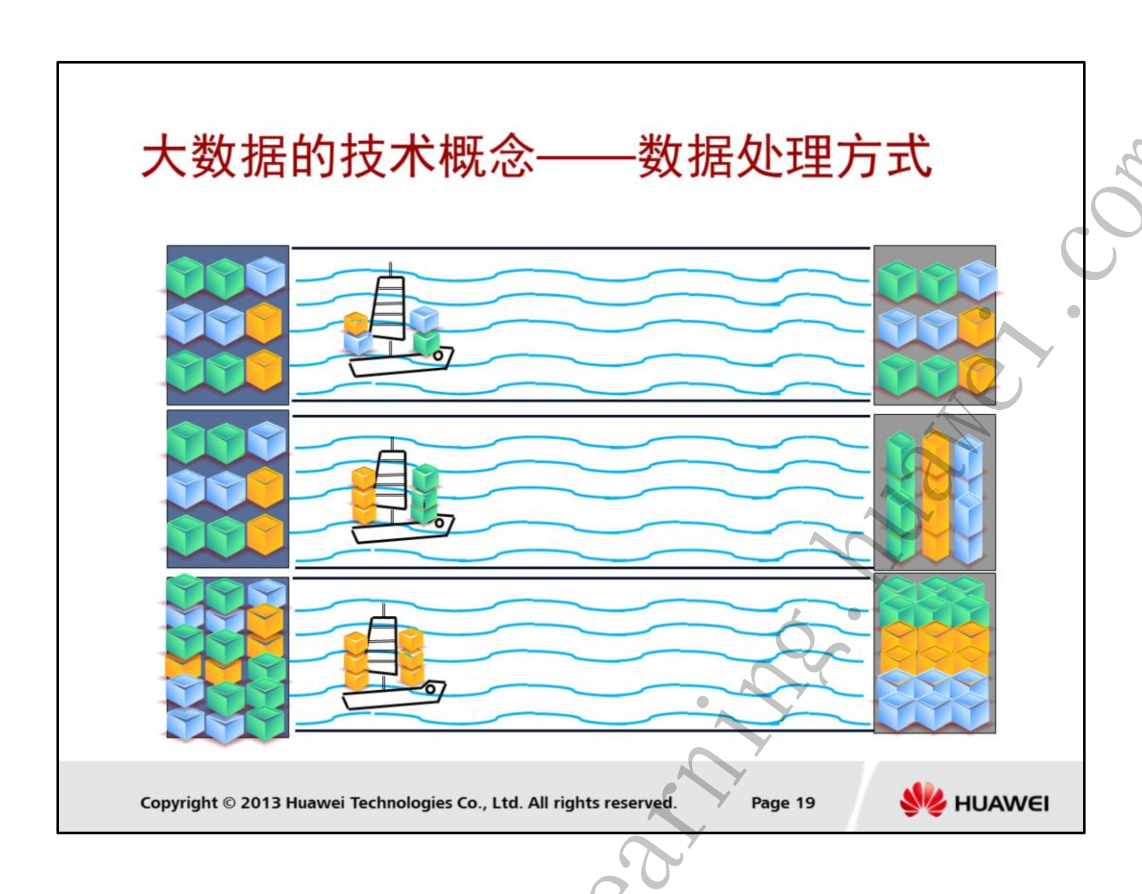
当只有少量的货物需要运输时，我们可以使用小船 ，当需要运输的货物数量增大时，就需要造大一些的船来运输，但当需要运输的货物数量变得很大时，我们就不能通过造更大的船来运输，因为河流的水深是一定，船太大反而无法通行。既然船不能做的无限大，而需要运输货物数量又越来越大，那么怎么来提高运载效率呢？大家知道，我们可以通过使用多条小一些的船分别运输来解决问题。需要运输的货物数量不断增加时，我们可以增加更多的小船，只是需要做好码头和航运的管理就可以了。

同样，在大数据时代，我们也不可能采用无限升级CPU、内存、磁盘容量等硬件能力来满足大数据处理的需求，而是通过采用新的方式来处理，即分布式并行处理。分布式并行处理方式的核心是充分利用了扩展性（类似于船运中河流的宽度），并发的进行数据处理。它不在是以往那种部件扩展（CPU 内存等）而是组件级的扩展。分布式并行处理扩展更灵活，承载能力更大，但是需要更强的调度和协调能力，也就是管理上变得复杂了，需要统一架构，统一管理。当然这个复杂不是对外的而是指对编写程序的复杂度和代码的执行效率要求更高了。

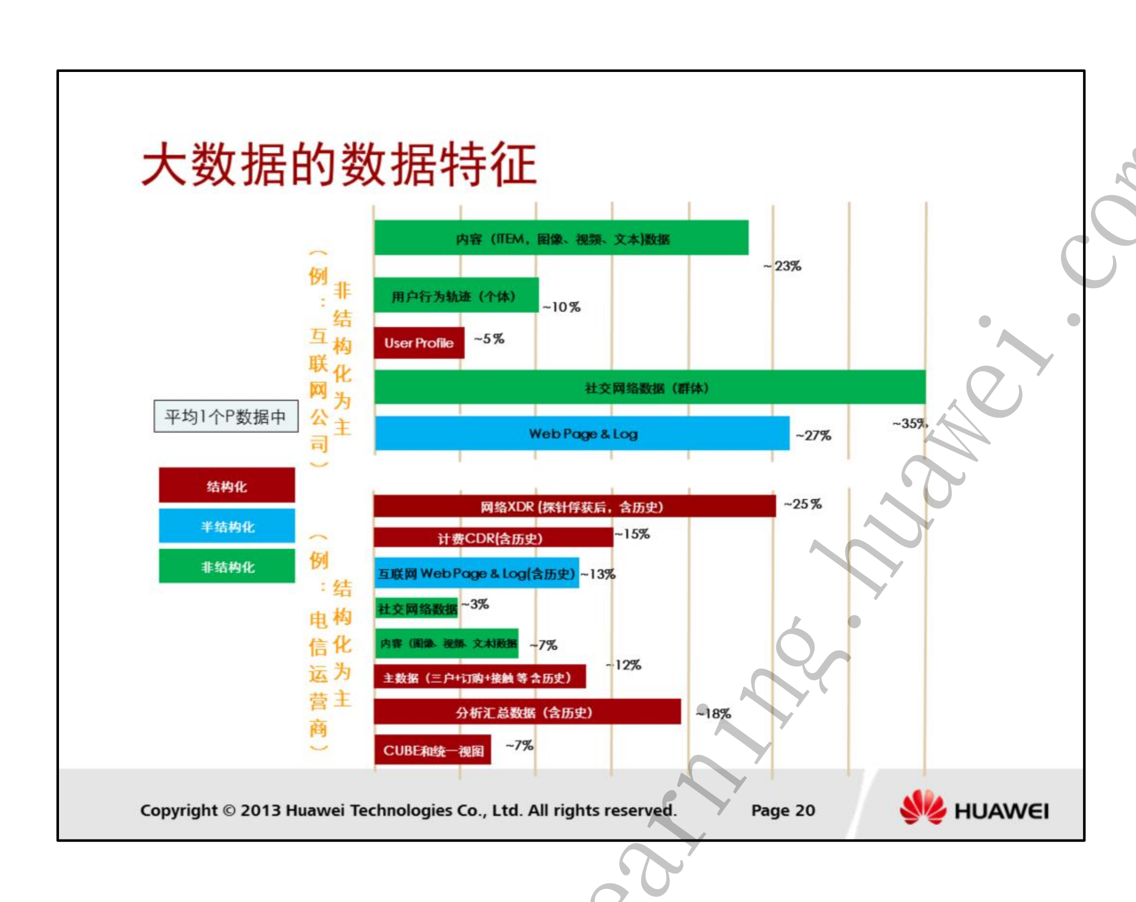


我们在大数据的4V特征中介绍了其中一个V——VALUE价值密度，这个价值密度是指数据的价值比例。有些数据价值密度会低一些，单看每一条感觉用处都不大，但从大量的低价值数据中整合分析可以获得高价值的结果（比如互联网数据）。对于企业的一些数据，尤其是数据库形成的数据，它的每一条都很重要，针对这些不同类型的大数据，应采用不同的处理手段。

如图示中，廉价货物使用普通船运输。同理，价值密度低数据对存储无特殊要求，通常采用不带任何增值特性的廉价存储。中高价值货物使用专业船来运输，保证货物的安全和运输效率。同样，在大数据处理中，高价值密度数据对存储有较高要求，需要专业含有多种保护机制及增值特性的存储，从而保证数据存储的安全性和高性能。



我们看到对不同价值数据处理会有不同处理方式，不仅体现在存储技术的选择上（船），在数据管理上也会有所不同。在第一艘船中，数据在加载过程中不做任何处理，直接传递给分析工具，由分析工具进行数据处理，那么分析工具据需要承担大量的数据筛选工作。第二艘中，数据做了粗加工处理，然后传递给分析工具，分析工具减少负担，会更快的给出结果。第三艘船中，数据在运载过程中做了分门别类的整理、细加工处理，然后传递给分析工具，分析工具直接导入数据，给出结果，分析工具不需要承担额外的消耗，专心做分析处理。



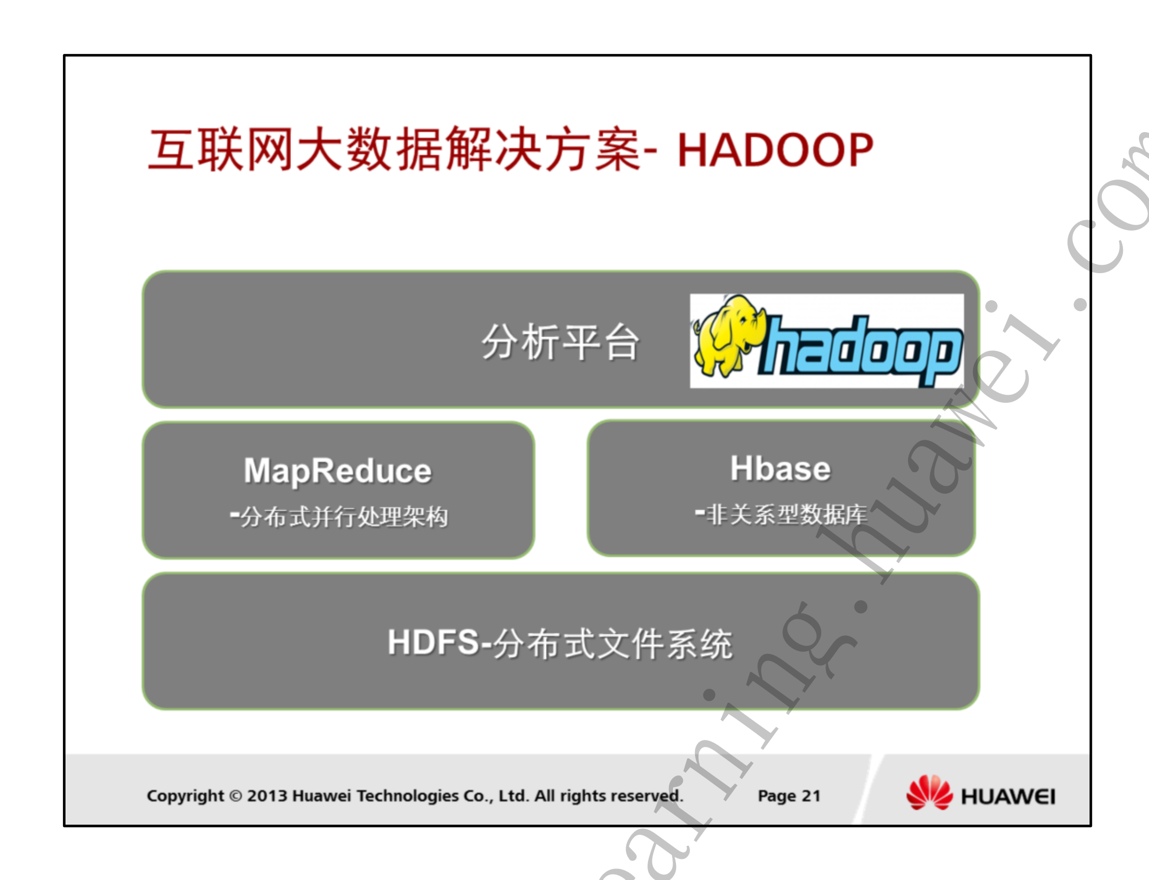
从数据结构特征来分类，主要可以分为：结构化数据、非结构化数据和半结构化数据。

结构化数据即指可以用二维表结构来逻辑表达实现的数据，简单来说就是数据库 。比如企业ERP、财务系统、医疗HIS数据库、教育一卡通、政府行政审批等。

非结构化数据，即不方便用数据库二维逻辑表来表现的数据。包括所有格式的办公文档、文本、图片、XML、HTML、各类报表、图像和音频/视频信息等等。比如医疗影像系统、教育视频点播、视频监控、国土GIS、设计院、文件服务器（PDM/FTP）、媒体资源管理等。

半结构化数据，包括邮件、HTML、报表、资源库等等，典型场景如邮件系统、WEB集群、教学资源库、数据挖掘系统、档案系统等等。

在这里我们主要分析两类大数据应用场景，一类是以互联网为代表，其90%以上数据都是非结构化数据，如Google，百度Facebook，Twitter，新浪等，数据以非结构化为主。另一类以运营商为代表，其80%以上是结构化为主的大数据。针对这两种大数据代表形式，在业界会有不同的技术处理方式。



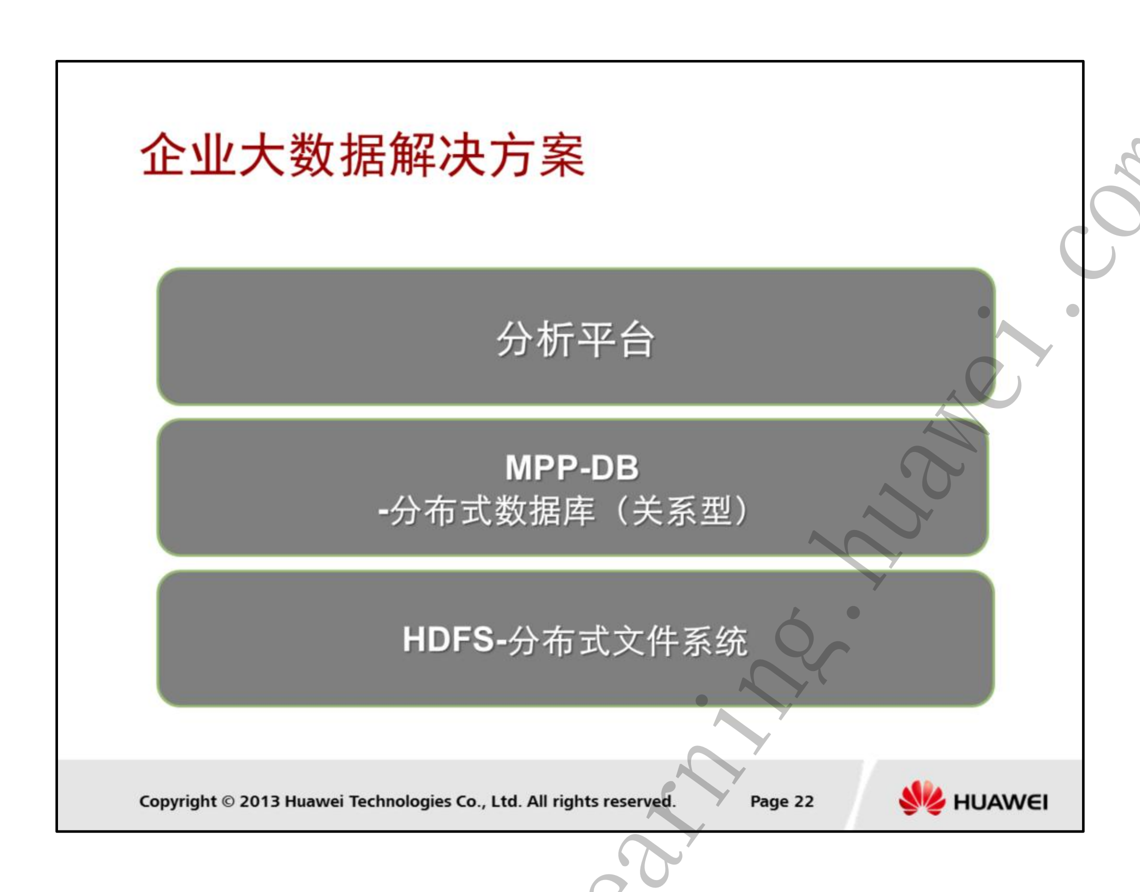
针对以非结构化数据为主的互联网大数据技术处理技术比较有代表性的是Hadoop。Hadoop是一种开源的对大规模数据进行分布式处理的技术框架，在处理大数据中非结构化数据上有着性能和成本方面的优势。

Hadoop主要分为三个部分，分别是：HDFS（Hadoop Distributed File System）分布式文件系统、Hbase（Hadoop Database）非关系型数据库和MapReduce分布式并行处理架构。

HDFS有着高容错性的特点，并且设计用来部署在低廉的（low-cost）硬件上，用来在各个计算节点上存储数据，并提供了对数据读写的高吞吐率。 HDFS使用的是廉价的存储，对存储无特殊要求，可以是服务器上的存储，也可以是专业存储。它不含有任何数据保护的增值特性，每次存储三份数据，用多份数据存储来保证数据的可靠性，对底层硬件的可靠性要求不高。

Hbase（Hadoop Database）是非关系型数据库的一种，是构建在HDFS之上的分布式、面向列的存储系统。它具有高可靠、高性能、面向列和可伸缩的特性。HBase适合于存储大表数据（表的规模可以达到数十亿行以及数百万列），并且对大表数据的读、写访问可以达到实时级别。

MapReduce是Google提出的一种简化并行计算的编程模型，名字源于该模型中的两项核心操作：Map和Reduce。Map将一个任务分解成为多个任务，Reduce将分解后多任务处理的结果汇总起来，得出最终的分析结果。最具有代表性的开源实现是Apache的Hadoop MapReduce。



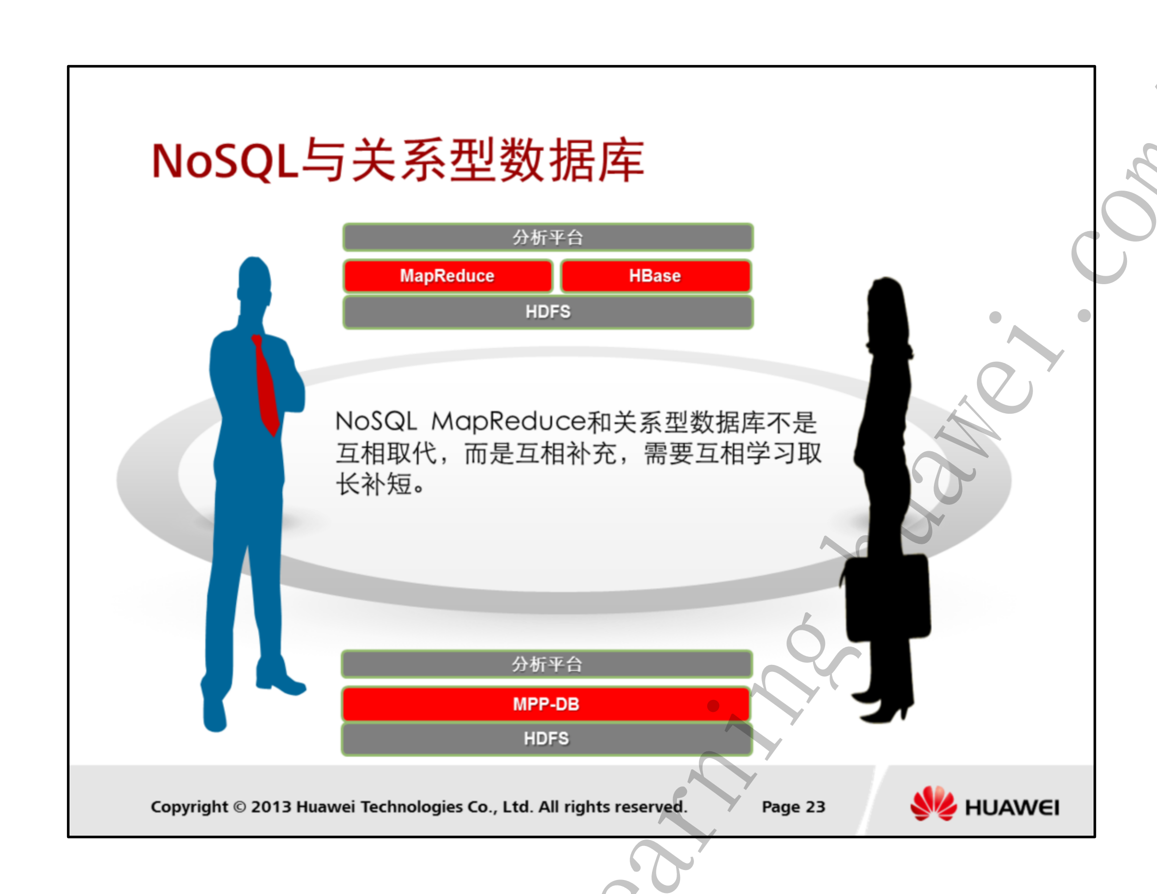
针对企业中结构化数据的大数据解决方案，在架构上与非结构化数据解决方案差别体现在数据库管理上。

相比于互联网大量低价值数据而言，企业数据价值远远高出很多，因此需要更可靠、更高性能的存储支撑。前面介绍的HADOOP采用存储三份数据，用多份数据存储来保证数据的可靠性，没有任何软件保护机制，硬件失效是常态的处理方式并不适合企业应用。

传统关系型数据库使用时间长，配套工具多，稳定性可靠性高。但是由于层级关系复杂，导致分支越多处理时间消耗越多,因此扩展能力弱。难以处理TB级数据，不能很好的支持高级别的数据分析。

并行数据库系统（Parallel Database System）是新一代高性能的数据库系统，其原理是把本来很复杂的关系分解到各个独立单元中区，这些单元彼此封闭，单个单元中关系级层少，这是分布型关系数据库的核心。把大库变成多个小库分布在不同的节点上，并行计算一个单元损坏不影响其它单元的运作，而且继承了关系型数据的的所有的优势。

利用并行数据库，我们还可以在运载和存储数据过程中对数据做更细的分类，在需要进行数据分析处理时，分工具BI不需要再做数据分类处理，直接进行数据分析得出分析结果，大大提升了数据分析处理的效率。

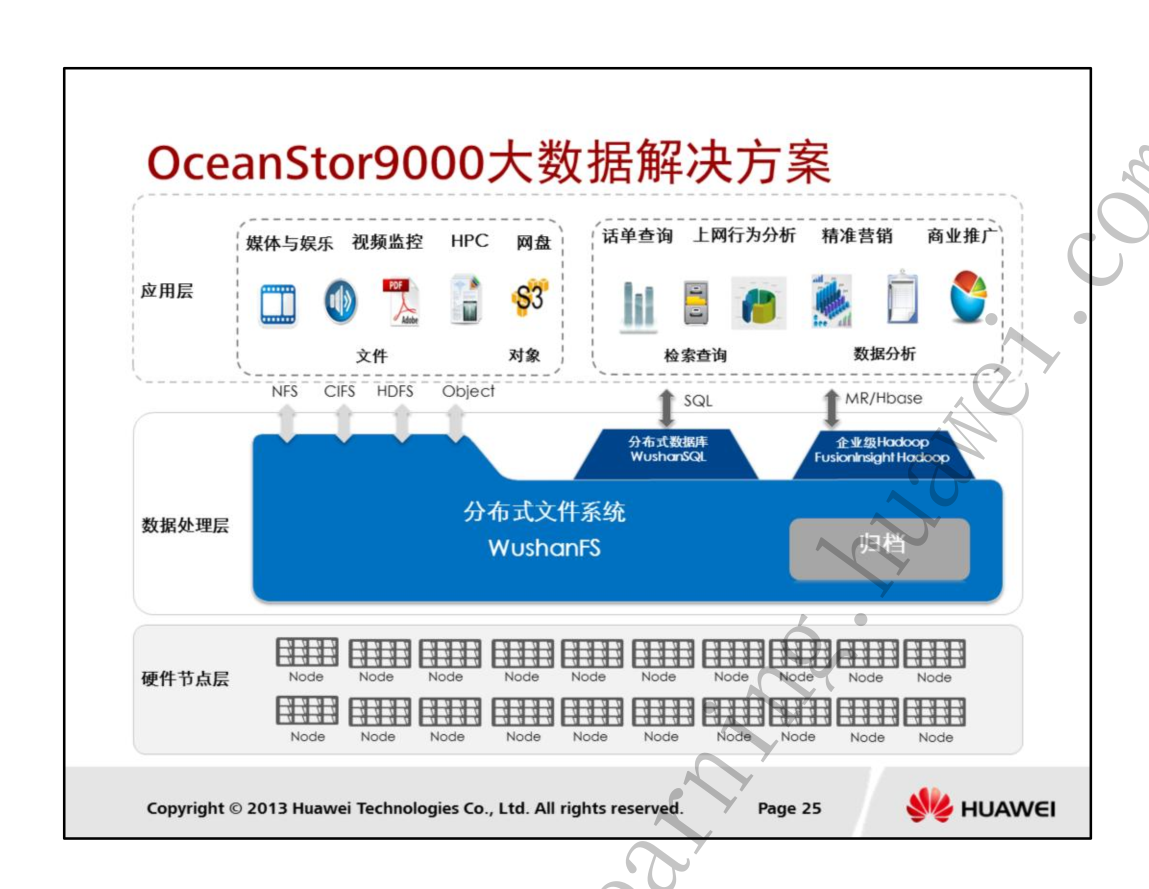


NoSQL，泛指非关系型的数据库,在应对Web2.0超大规模和高并发的动态网站时，比传统关系型数据库更灵活，大数据量高性能，易扩展。NoSQL　MapReduce和关系型数据库不是互相取代，而是互相补充，需要互相学习取长补短。



目录

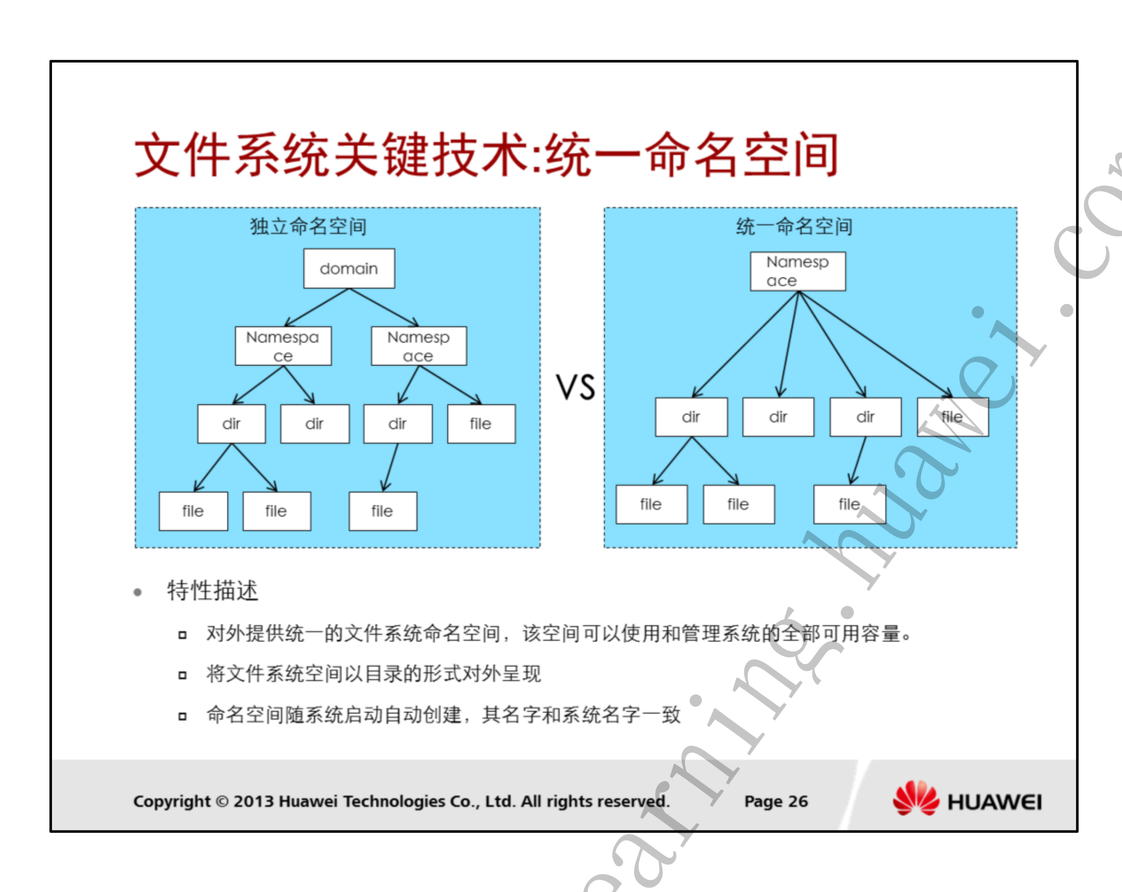




华为OceanStor 9000大数据解决方案，集大数据的存储、备份、分析为一体化（统一管理、统一硬件平台、同一组网），极大简化组网和管理复杂度；文件系统直接管理底层磁盘，减少了复杂的RAID配置、LUN划分等步骤；

所有的节点都统一于OceanStor 9000硬件平台，内部网络适用10GE或者是Infiniband High-speed，因此OceanStor 9000可以在保障低延时、高带宽、高并发的同时表现极好的性能。为了匹配不同的应用场景，OceanStor 9000的节点分为高OPS节点、高带宽节点和大容量节点，用户可以根据不同的商业性能和容量的需求，灵活的配置不同节点的数量。

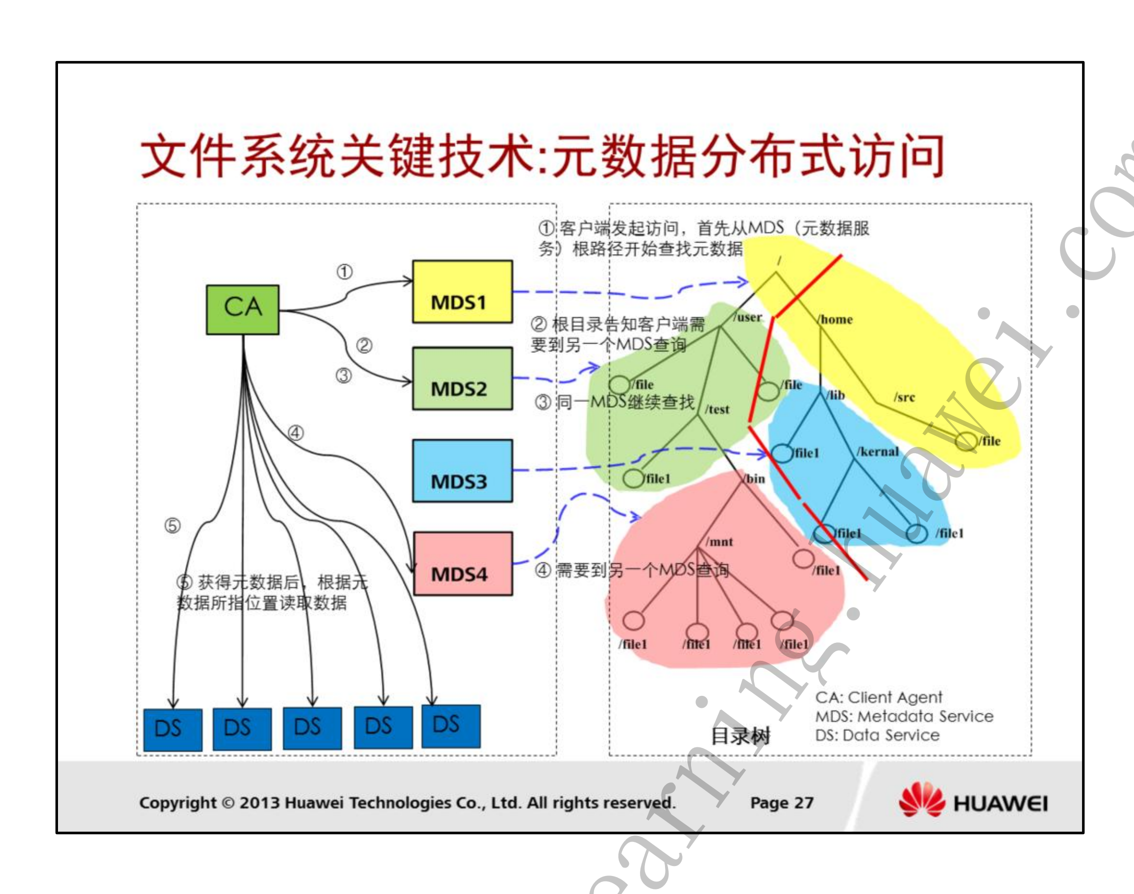
OceanStor 9000支持多种接口与数据类型，如NAS接口NFS、CIFS和POSIX；目标接口REST和SOAP,数据库接口JDBC和ODBC；备份归档接口VTL,OST等。应用这一特性，OceanStor 9000解决方案完全可以胜任核心生产数据存储，商业数据存储与分析。



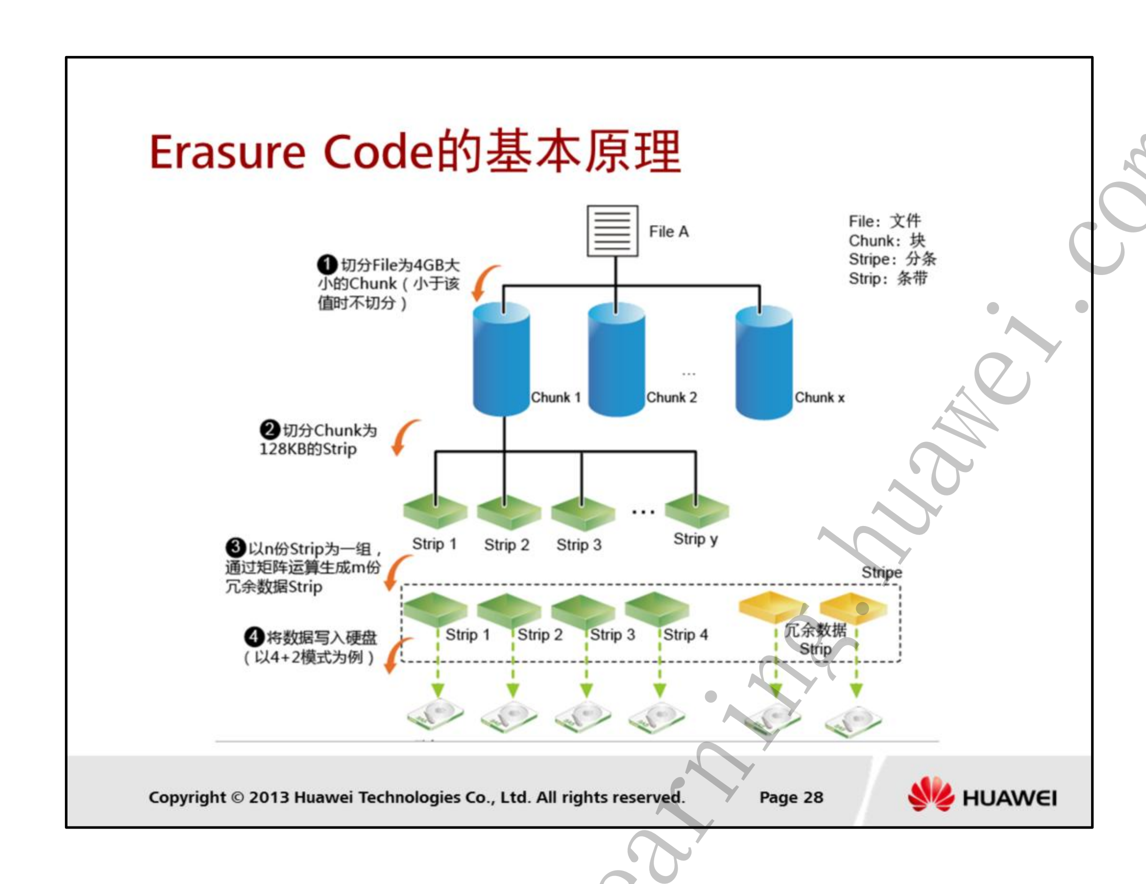
WushanFS为运行于Linux操作系统之上的分布式文件系统，让存储节点集群统一对外提供CIFS与NFS文件共享服务。

Wushan分布式文件系采用全对称结构，将所有节点的命名空间整合为统一命名空间，将整个系统的所有节点的存储容量组成一个大的虚拟存储池，元数据和数据保存在每个节点上，每个节点都是元数据服务器同时也是数据服务器，访问文件数据时wushan文件系统首先定位到文件所归属的元数据服务器，然后通过元数据服务器获取文件的数据分布，即获取文件分布在哪些节点上以及在节点上的具体位置，再去访问这些节点完成数据读写操作。

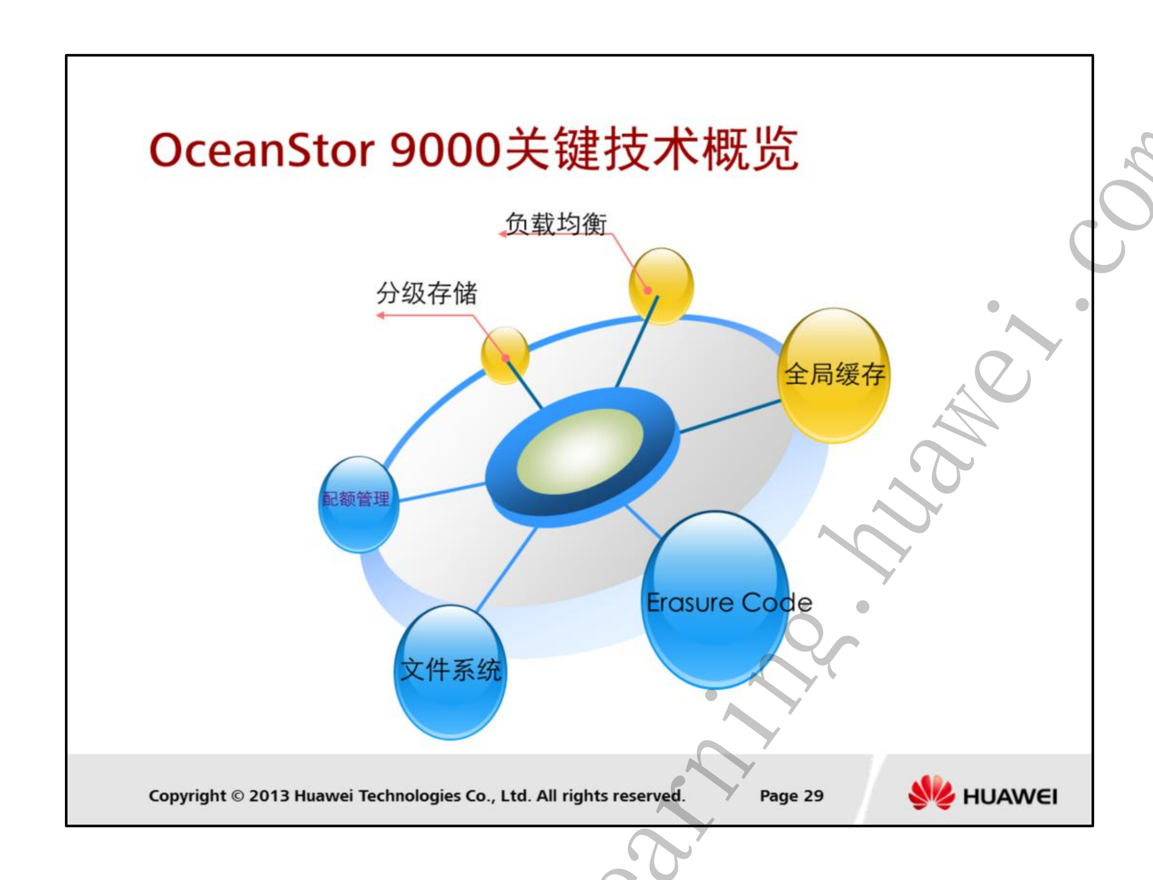
- 客户价值
  - 为用户提供统一的存储资源池，不论系统规模多大，都可以集中进行管理和分配
  - 系统扩容时，扩充的容量可直接被纳入资源池使用，避免维护多套存储设备的困难。



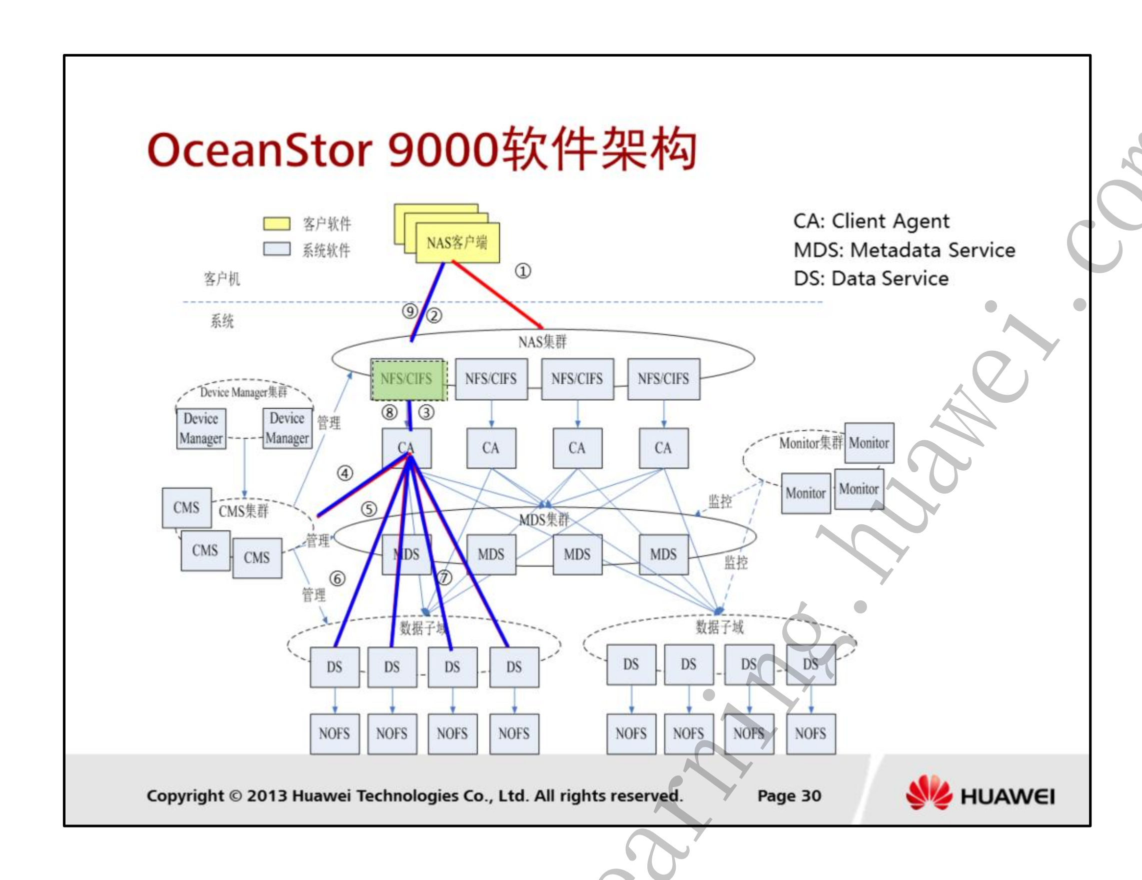
- 使用动态子树方式组织元数据
- 将命名空间中的所有元数据按照名字子树的方式划分，每棵名字子树划分给一个MDS处理。一个MDS可以管理若干棵子树。
- 正常处理流程：CA将报文发给任意一个MDS，该MDS查询本地动态子树表，如果属于本地MDS处理范围，则直接处理。如果不属于则将该请求转交给负责该名字子树的MDS（如果无法直接找到，则转给最可能负责该名字子树的MDS）
- 故障迁移：一台MDS故障后，其他MDS将分担故障MDS负责管理的名字子树。
- 负载均衡：当MDS数组增减或某个MDS负责的名字子树访问过冷/过热时，会采用负载均衡策略，将过热节点上的名字子树部分迁移到较冷节点管理。



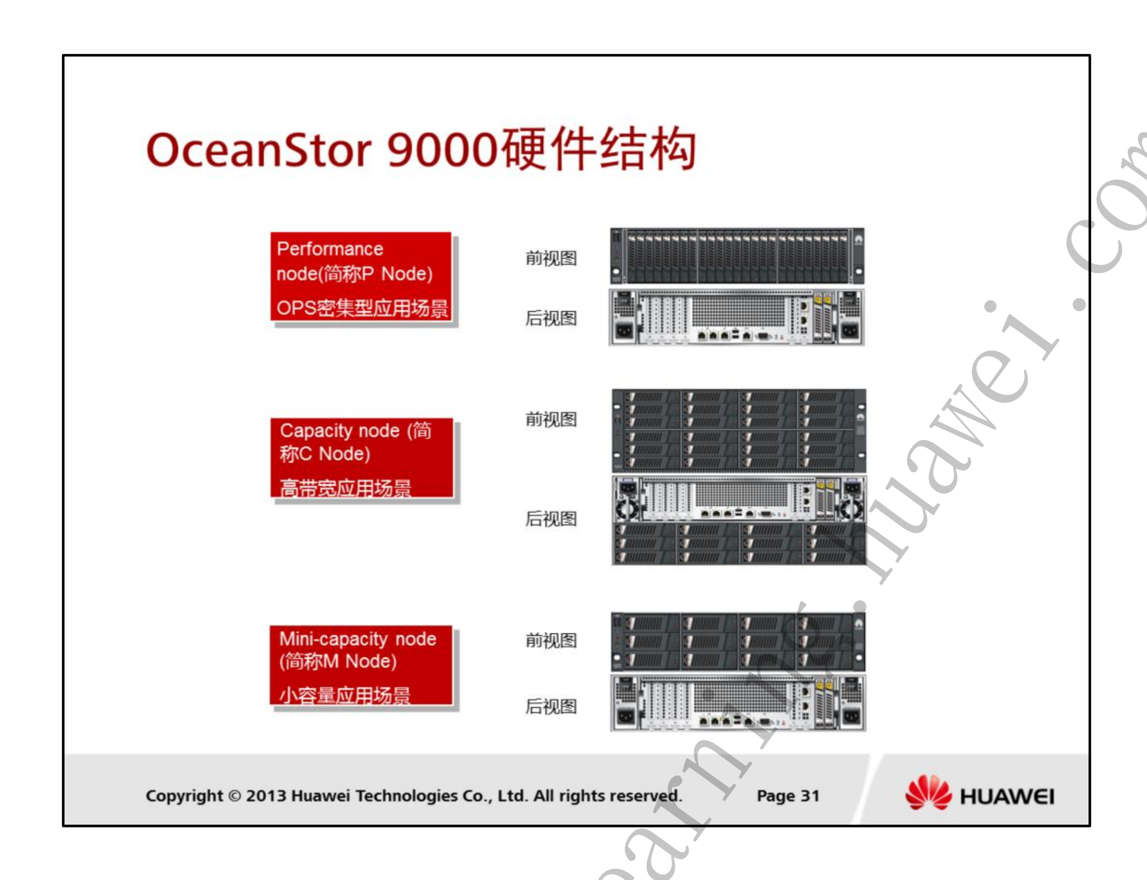
- Strip大小缺省为128KB,当数据不足128KB时，Strip可以为16KB的整数倍。
- 当预期一个目录内平均文件大小小于512KB时，为提升硬盘利用率，管理员可手工指定该目录内所有文件采用的Strip大小为16KB。



- 负载均衡
  - 负载均衡特性，提供接入负载均衡、域名访问和节点故障后业务无中断切换功能。
- 分级存储
  - 自动分级存储特性，可实现热数据与冷数据在不同存储级别间的智能存储和动态迁移。
- 全局缓存
  - OceanStor 9000通过全局缓存模式，整合所有节点缓存，同一文件的数据在缓存池只缓存一份，任意节点均可命中，有效提高数据访问命中率，减少硬盘读写次数，降低访问时延，提升系统整体性能。
- 文件系统
  - WushanFS为运行于Linux操作系统之上的分布式文件系统，让存储节点集群统一对外提供CIFS与NFS文件共享服务。
- Erasure Code
  - OceanStor 9000采用Erasure Code算法，提供更多的数据冗余，由于采用数据切片的技术，在数据恢复时效率更高。
- 配额管理
  - OceanStor 9000提供的配额类型包括
    - 容量配额：管理和监控存储空间的使用情况。
    - 文件数配额：管理和监控文件数量的使用情况。



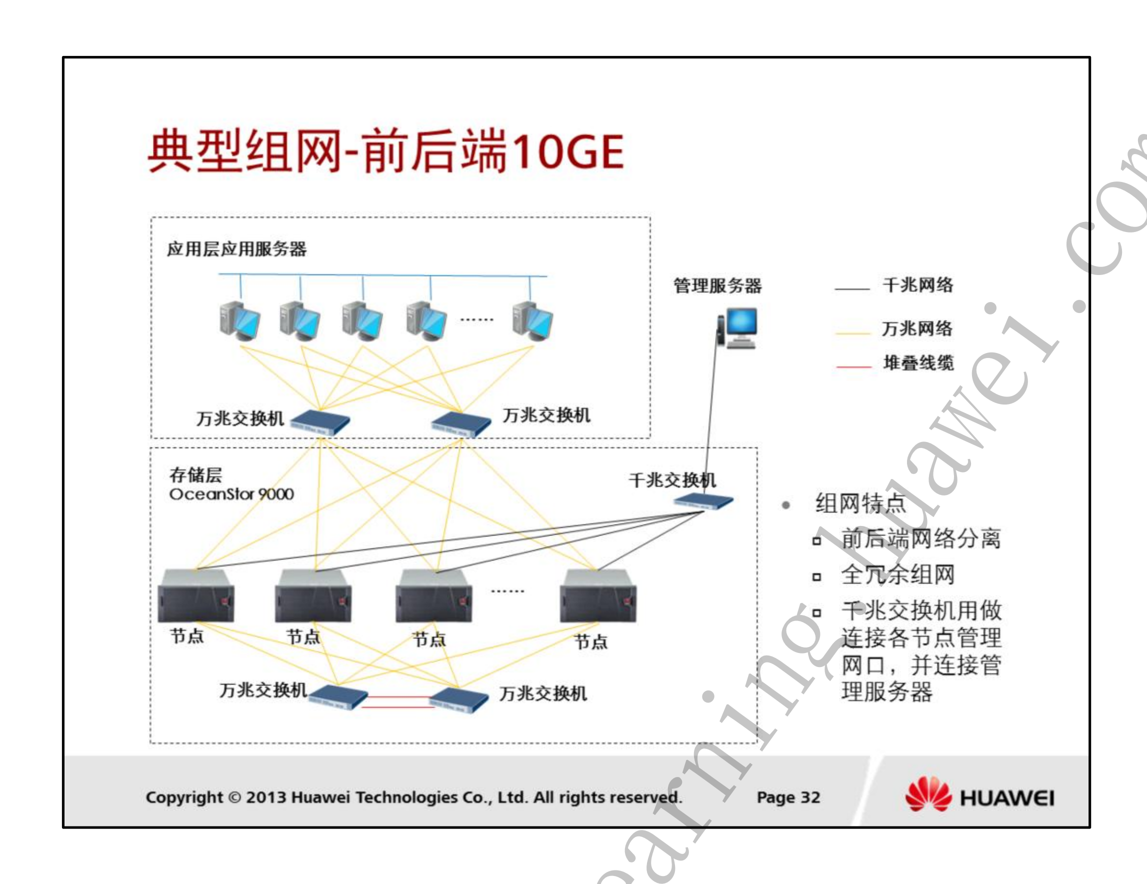
- 访问协议集群域名
- 和服务端建立连接/发起业务
- 业务下发任务给CA（OceanStor 9000内部访问客户端）
- CA向MDS（元数据服务）查询元数据
- 返回查询结果(图上画了一次，实际可能有多次交互)
- CA从多个节点读取数据和校验数据
- 返回数据
- CA向NAS服务器返回结果
- NAS服务器向客户端返回结果



OceanStor 9000的硬件系统主要由存储节点、网络设备、KVM和调制解调器组成。

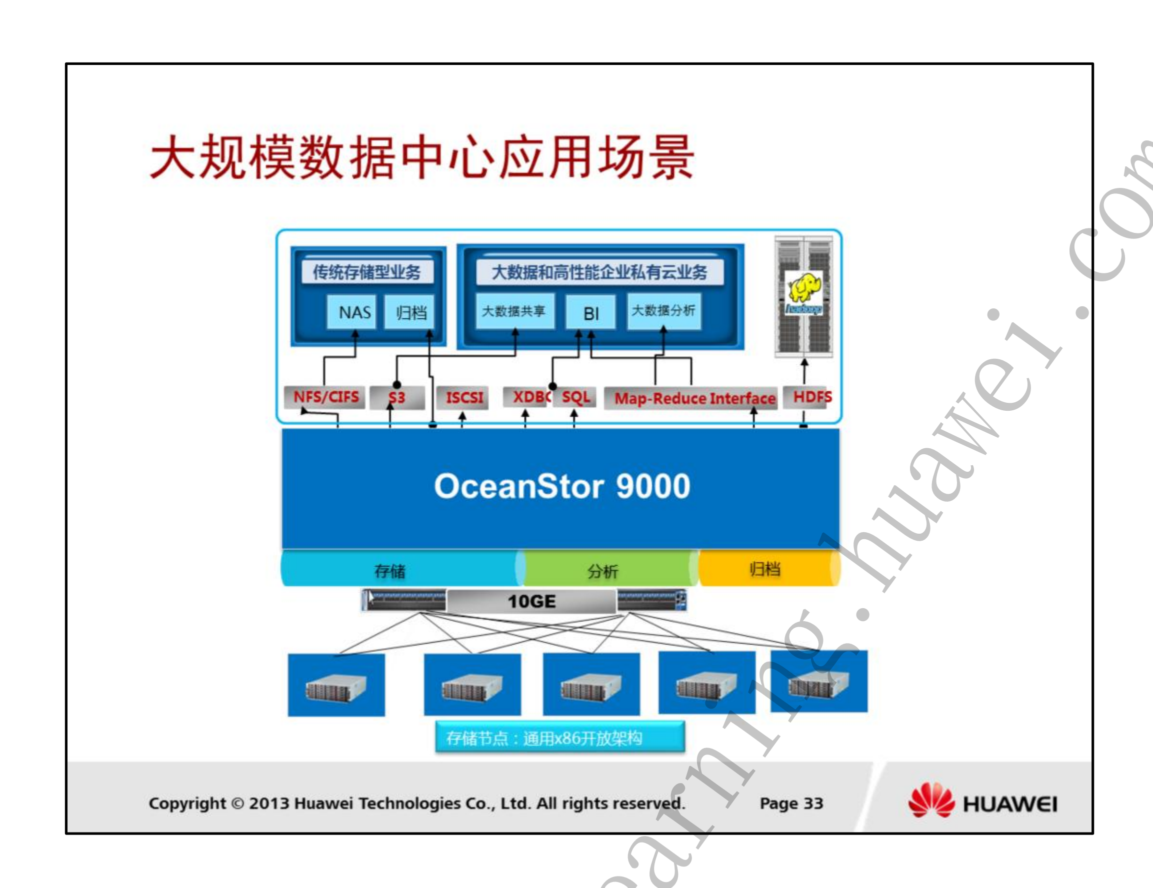
存储节点可选用P节点,C节点和M节点，P节点适用OPS密集型应用的场景，C节点适用高带宽应用场景,M节点适用于小容量应用场景。

- P节点
  - 2U 25盘（2.5寸数据盘）设备
  - 标配：2路6核CPU，48G内存，
  - 4*200G SSD+21*600G SAS盘
- C节点
  - 4U 36盘（3.5寸数据盘）设备
  - 标配：2路6核CPU，48G内存，1*200G SSD+35*4T SATA盘
- M节点
  - 2U 12盘（3.5寸数据盘）设备
  - 标配：2路4核CPU，32G内存， 1*200G SSD+11*2T SATA盘



- OceanStor 9000的组网结构包括前端业务网络和后端存储网络。
- OceanStor 9000的组网结构包括：
  - 前端业务网络用于OceanStor 9000与用户网络对接；
  - 后端存储网络用于OceanStor 9000内部节点间互联；
  - IPMI网络用于OceanStor 9000与用户维护网络对接。

OceanStor 9000支持多种组网方式，包括10GE组网、IB组网和GE组网等，可满足用户的不同组网需求。



- 高性能
  - 全球领先的读写性能，性能超过500万+，可以很好的满足高性能应用的需求
- 数据快速迁移
  - 存储到分析数据快速装载，智能流动
  - 快速查询: 最快一秒钟可检索100亿条记录
- 数据智能生态链
  - 数据存储、分析、归档一体的全生命周期管理

总结

- 大数据的定义与特性
- 对象存储技术
- 大数据处理技术
- 华为OceanStor 9000大数据产品架构与特性



思考题

1. 大数据与传统数据的主要区别是什么？
2. 对象存储系统由哪些部分组成？
3. 传统关系型数据库为什么不适用于大数据？
4. 华为大数据产品是什么，有哪些特性？



练习题

- 判断题

  1、 Hbase是NoSQL。（　　）

  2、HDFS不含有任何数据保护的增值特性，每次存储三份数据，用多份数据存储来保证数据的可靠性，对底层硬件的可靠性要求不高。（　）

  3、 NoSQL MapReduce将取代关系型数据库（　）





练习题

- 多选题

1、大数据有哪些特性？（ ）

A、Volume

B、Variety

C、Value

D、Velocity

2、华为OceanStor 9000大数据解决方案的主要特点包括？（ ）

A、集大数据的存储、备份、分析为一体化

B、保障低延时、高带宽、高并发的同时表现极好的性能

C、支持多种接口与数据类型

D、分为高OPS节点、高带宽节点，用户可以根据不同的商业性能和容量的需求，灵活的配置不同节点的数量

 Page 37



- 习题答案
  - 判断题：1、T 2、T 3、F
  - 多选题：1、ABCD 2、ABCD

Thank you

www.huawei.com

HC1109109 备份容灾技术基础




HC1109109
备份容灾技术基础
www.huawei.com
Copyright © 2013 Huawei Technologies Co., Ltd. All rights reserved.
HUAWEI



目标

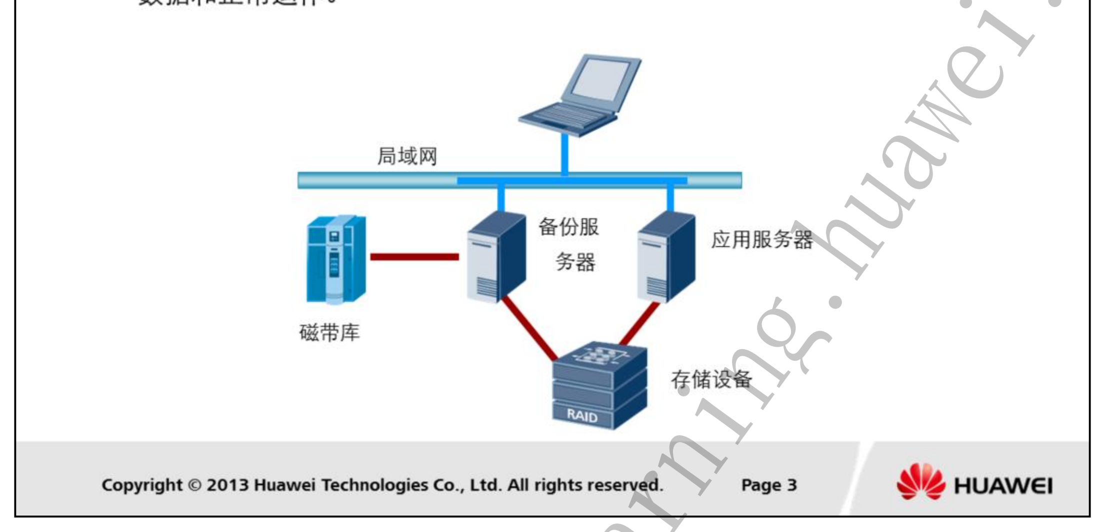
- 学完本课程后，您将能够:
  - 了解备份概念及拓扑结构
  - 理解备份技术
  - 掌握备份策略制定
  - 熟悉华为备份实现与应用
  - 了解容灾





目录





在一套备份系统中，通常包括以下组件：

- 备份服务器

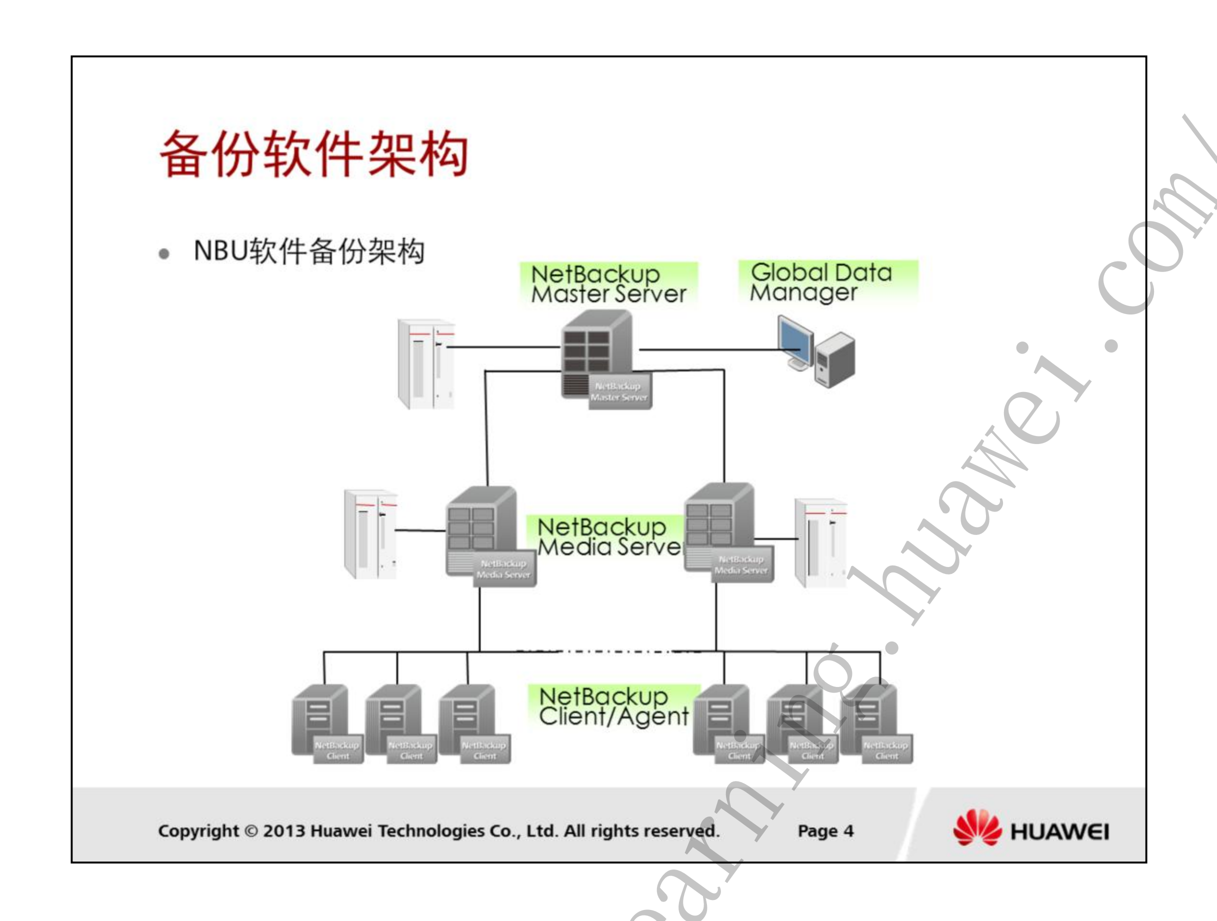
备份服务器是运行备份软件的载体，一般是PC服务器和小型机。

- 备份软件

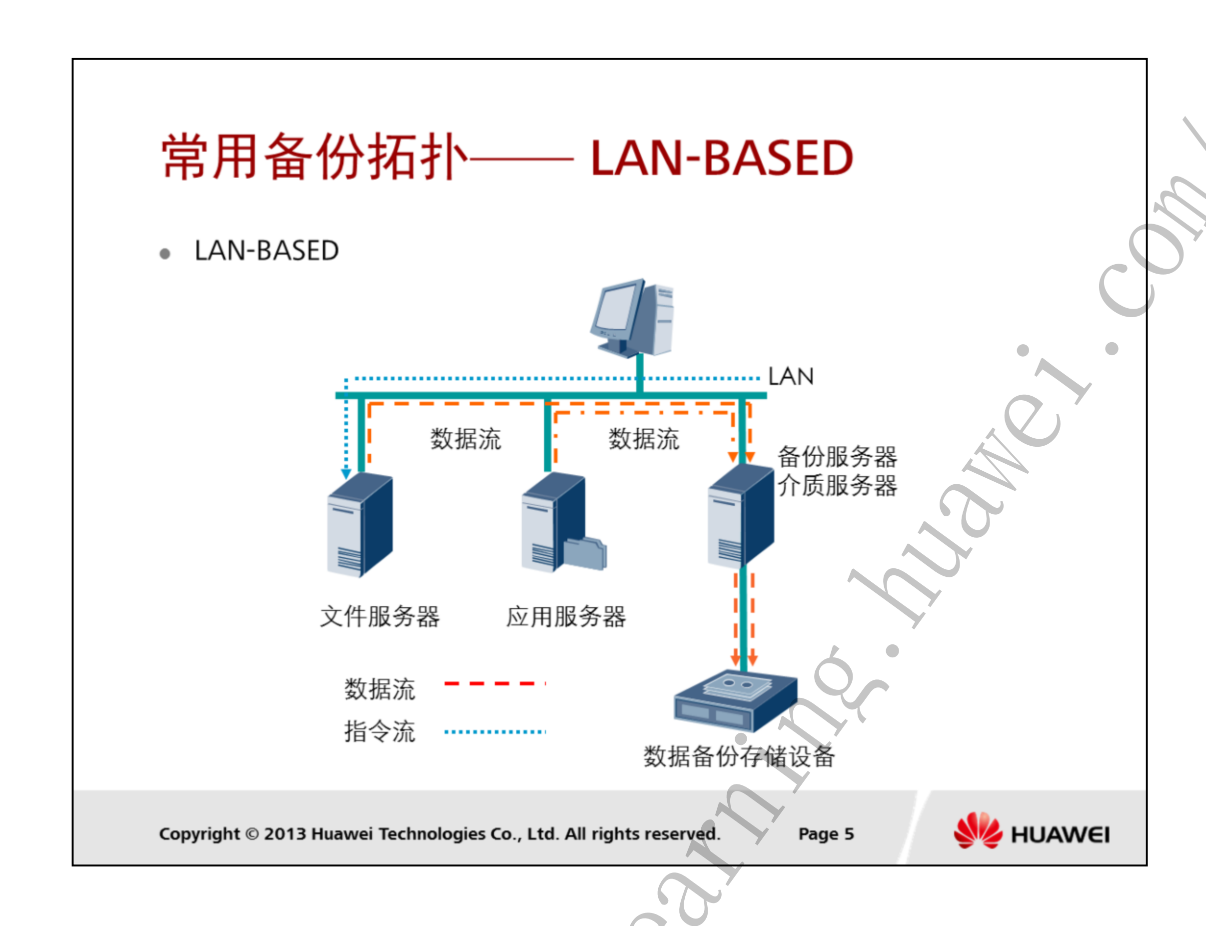
备份系统的核心，备份软件控制生产数据拷贝到存储介质上，并对备份数据进行管理，如Symantec的Backup Exec和NetBackup等。

- 存储设备

备份数据存储的设备，一般为磁盘阵列、物理磁带库或者虚拟带库。



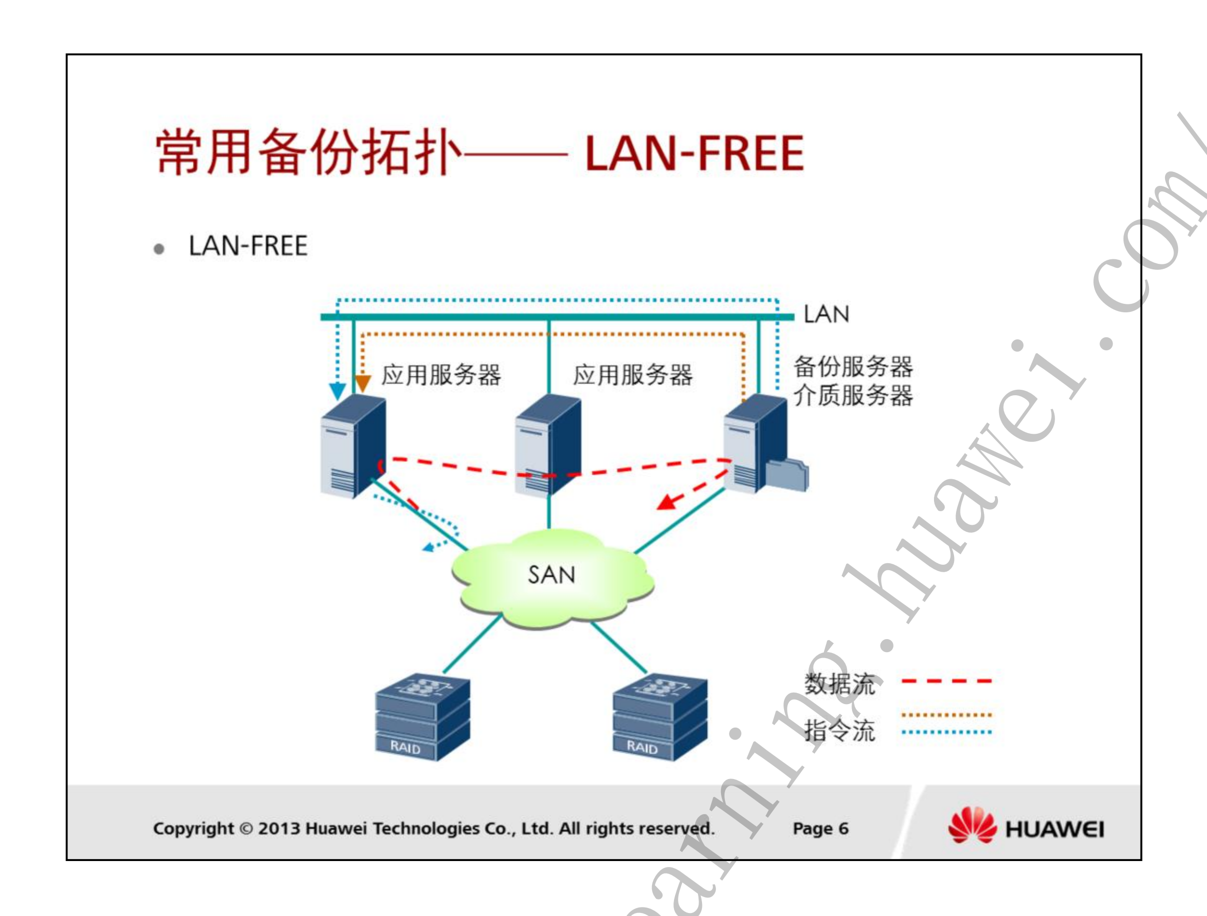
- NBU备份软件组件主要有：
  - Master Server(管理服务器)对备份系统的个模块的系统管理，对备份策略、备份任务以及数据恢复的配置界面和过程监控。
  - Media Server（介质服务器）负责介质设备的管理，和介质设备通信、读写管理，是备份服务器和介质的中间件。
  - Client（备份用户端）备份目标设备，软件模块client负责和Master server通信交换信息。
  - Agent（数据库代理）只在数据库备份时需要。
  - Management Console（管理终端）是用户GUI，可视化操作界面，管理备份软件操作。



基于LAN的备份，数据流和控制流都基于LAN传输，占用网络资源的备份模式。

备份数据流向：备份服务器通过LAN发控制流到安装了代理的应用服务器上，应用服务器响应请求通过LAN发送数据到备份服务器，备份服务器接收数据存储到存储设备上，完成备份。

- 优点：
  - 备份系统和应用系统分开，备份时不占用应用服务器的硬件资源。
- 缺点：
  - 增加额外的备份服务器硬件投资成本。
  - 备份的代理会影响应用服务器的性能。
  - 备份数据基于LAN，影响网络性能。
  - 需要单独维护备份业务，增加管理和维护的难度。
  - 对用户业务处理能力的要求较高。



顾名思义不占用LAN资源，控制流基于LAN传输，数据流不经过LAN。

备份数据流向：备份服务器通过LAN发控制流到安装了代理的应用服务器上，应用服务器响应请求读取生产数据，介质服务器直接从应用服务器上读取数据传输到备份介质上，完成备份。

- 优点：
  - 备份数据流不占用LAN资源，大大提高备份性能，且不影响用户网络性能。
- 缺点：
  - 备份的代理会影响应用服务器的性能。
  - 对用户业务处理能力的要求较高。



目录



备份系统的组成

备份软件

完成备份策略的制定、备份介质的管理以及一些其它扩展功能

备份介质

磁带库
磁盘阵列
虚拟磁带库
光盘塔/光盘库

备份服务器

用于安装备份软件的服务器，通过备份服务器预先制定的备份策略将数据备份到备份介质

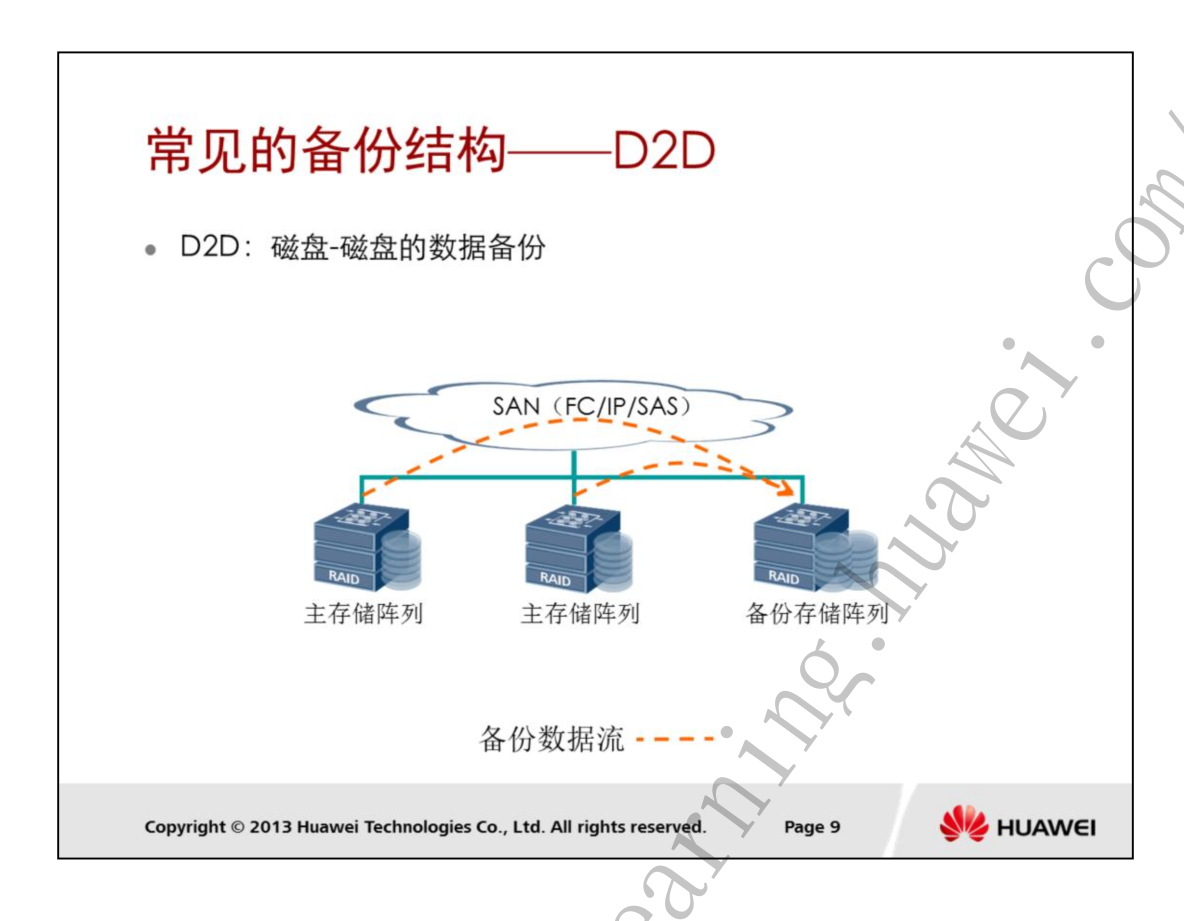
Page 8


完整的备份系统通常是由备份软件、备份介质和备份服务器构成的。

备份软件被用于制定具体的备份策略，同时对备份介质进行管理并应用备份介质进行数据的备份操作。通过备份软件的应用可以实现针对应用数据、应用程序甚至是应用系统的全面保护。一些扩展的备份软件能够实现更多的功能。完善的备份与恢复解决方案设计用于保护、备份、归档和恢复各类计算环境中的数据，这些环境包括大型公司数据中心、远程组、台式机及笔记本电脑。通过备份软件相集成可以提供数据整个生命周期内的管理解决方案 ——从创建到处理，从现场和非现场方式，跨越存储层次的所有级别 ，涵盖了磁盘、磁带和光学存储介质。备份软件的应用将能够做好更充分的准备，并且更加轻松地从设备故障、病毒攻击或意外丢失文件中恢复。

传统的备份介质通常是磁带库。除了传统的磁带库备份介质以外，还有基于磁盘的备份介质及虚拟磁带库（VTL）。

安装了备份软件的应用服务器就是备份服务器，备份服务器提供了备份软件运行的环境，并且为备份策略的执行提供服务。在某些备份架构中，备份服务器不仅提供到存储设备或者应用服务器的备份指令流，甚至也会接收应用发送过来的数据流并由其将数据送往备份介质进行统一、集中备份，所以备份服务器在备份架构中起着中枢的作用。

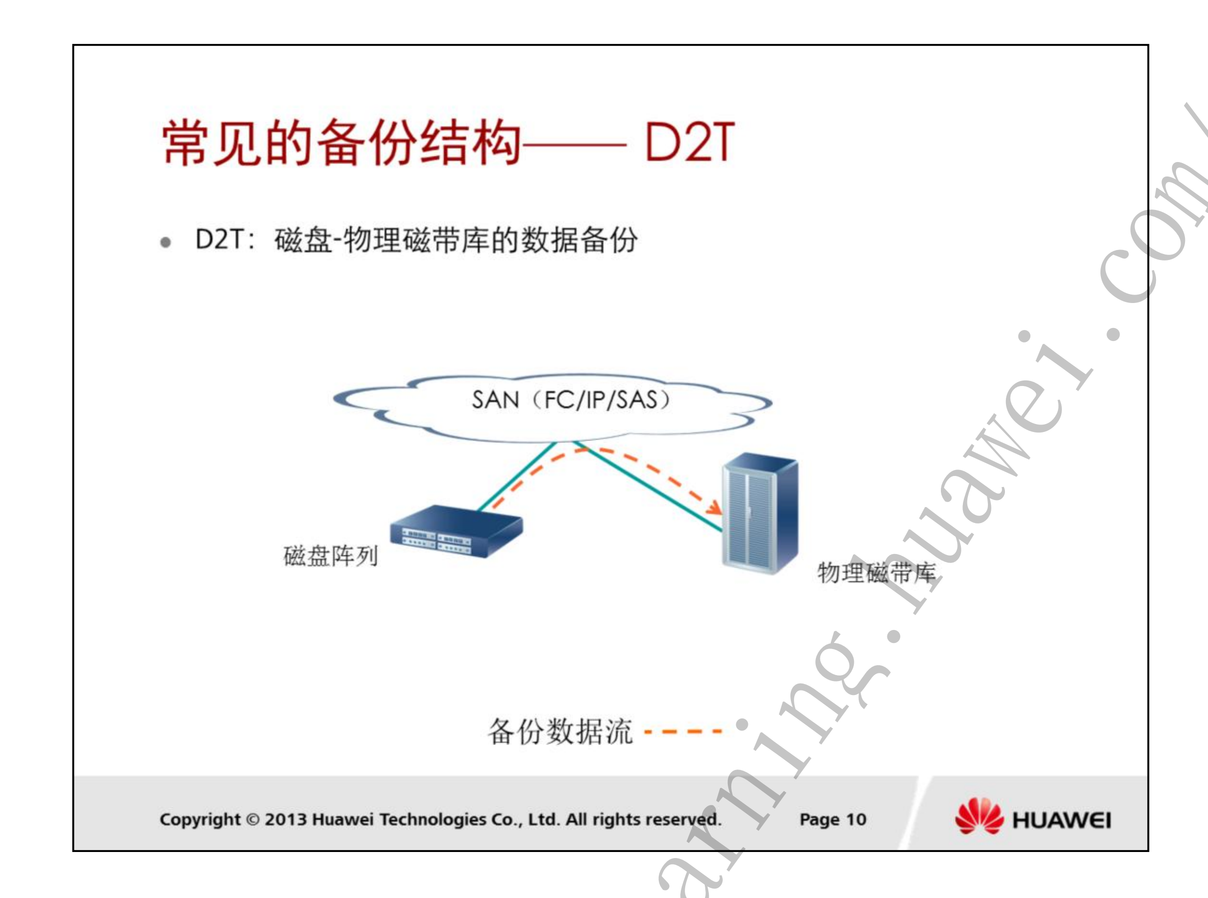


一个应用系统的数据备份决定了该系统的可靠性及可维护性，因此，数据备份系统的建设要充分考虑可靠性、可管理性及维护成本等几方面重要因素。目前常见的备份结构包括：

- 磁盘-磁盘的数据备份 (D2D) 。
- 磁盘-物理磁带库的数据备份 (D2T)。
- 磁盘-虚拟磁带库的数据备份 (D2V)。
- 磁盘-虚拟磁带库-物理磁带的数据备份(D2D2T)。

磁盘-磁盘(Disk to Disk，D2D)的备份是使用磁盘阵列作为主存储介质和备份存储介质的一种解决方案，具体有两种实现方式：

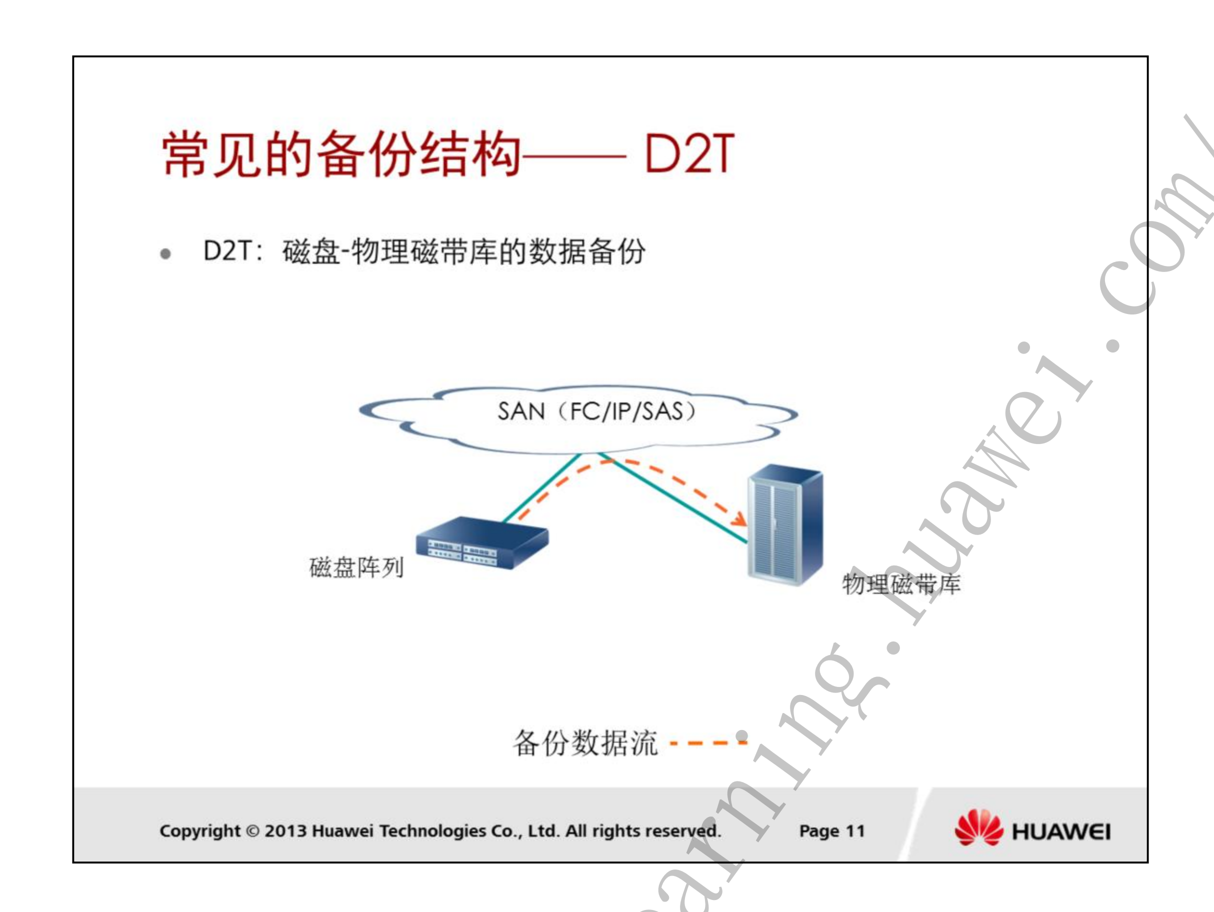
- 用户为备份系统部署一套磁盘阵列作为备份介质，通过备份软件将应用数据备份到备份服务器连接的磁盘阵列中。
- 用户为备份系统部署磁盘阵列作为备份介质，新部署的磁盘阵列与现有的在线存储磁盘阵列为同一品牌、同一型号，通过此类磁盘阵列所具备的LUN拷贝、快照或远程复制功能实现将现有磁盘阵列中的数据复制到备份磁盘阵列中的目的。



磁盘-物理磁带（Disk to Physical Tape Library,D2T）的备份是应用最为广泛的一种备份结构，有很多用户采用了这种备份结构对自身的应用数据进行备份；然而，在实际的使用中，很多用户发现，物理磁带库的故障使得整个备份系统的运行以及管理面临着很大的风险和挑战。

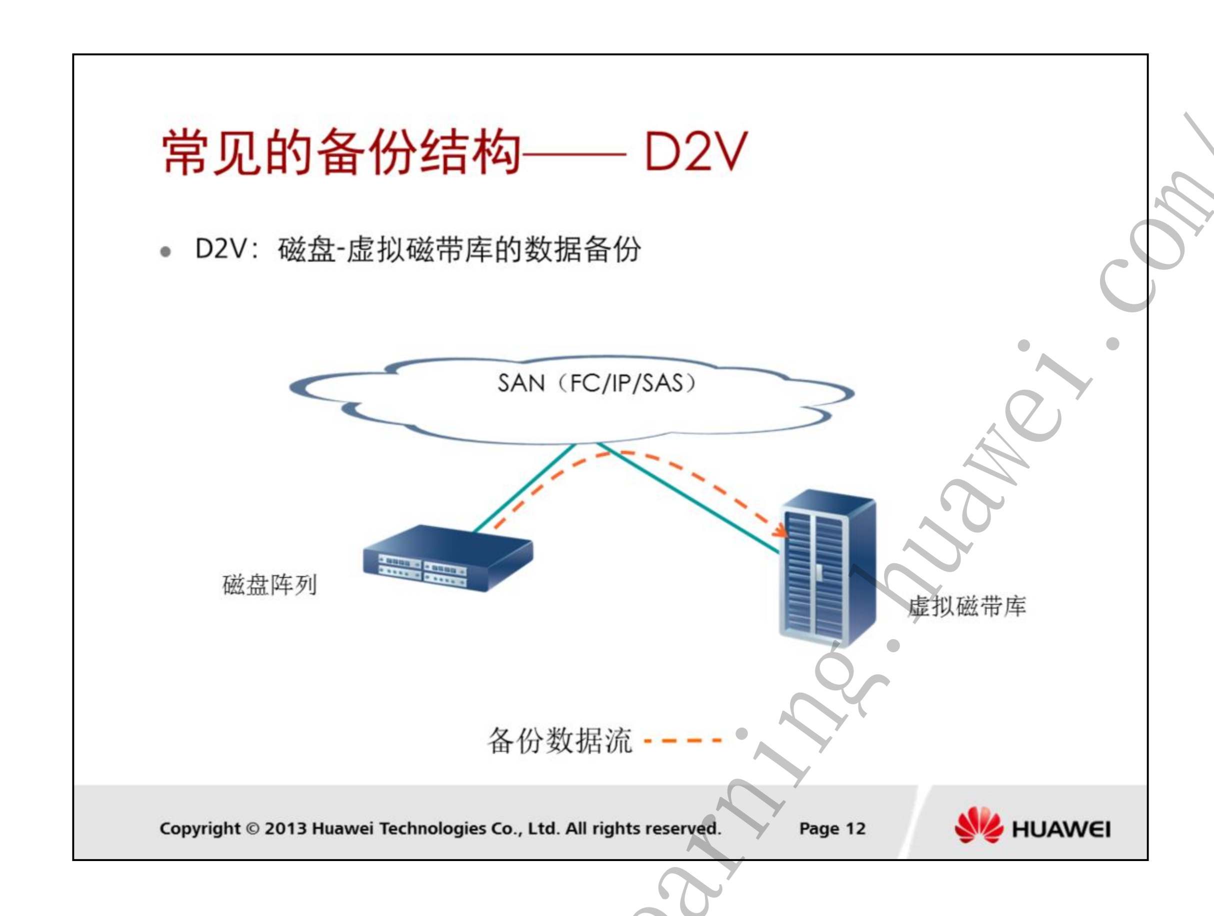
通过物理磁带库与备份软件结合可以方便的规划备份策略，然而，物理磁带库的故障却经常使得备份策略无法正确完成，甚至影响到整个系统的备份计划；IDG数据统计，用户平均每年用于维护物理磁带库的花费相当于部署该物理磁带库费用的15%，物理磁带库是由很多的精密机械元器件组成的一个更大的精密仪器系统，任何小的零部件磨损都有可能造成整个物理磁带库的宕机；目前来看，物理磁带驱动器以及磁带库的机械手臂是物理磁带库故障的最大原因，一旦物理磁带库出现故障，往往只能返回原厂或更换一台新设备，这个时间往往要持续几天甚至更长时间，在这个过程中，所有的备份作业都无法进行，备份策略会受到很大的影响。

物理磁带库的I/O瓶颈问题一直是困扰用户的另一个问题，由于物理磁带的读写是顺序进行的，无法像物理硬盘一样进行随机读写，这就决定了一个物理磁带驱动的I/O性能是固定的，如果现有的I/O性能无法满足要求，用户只能通过增加物理磁带驱动器的方式来提升性能。部署物理磁带驱动器的成本相对来说是很高的，并且，系统中每增加一个物理磁带驱动器，整个备份系统的稳定性就会增加一定的风险。



物理磁带库的物理磁带的可靠性会随着磁带使用次数的增多而不断下降，在使用物理磁带库作为备份介质的用户中，很多都经历过物理磁带损坏或无法读取而导致数据无法恢复的事故。

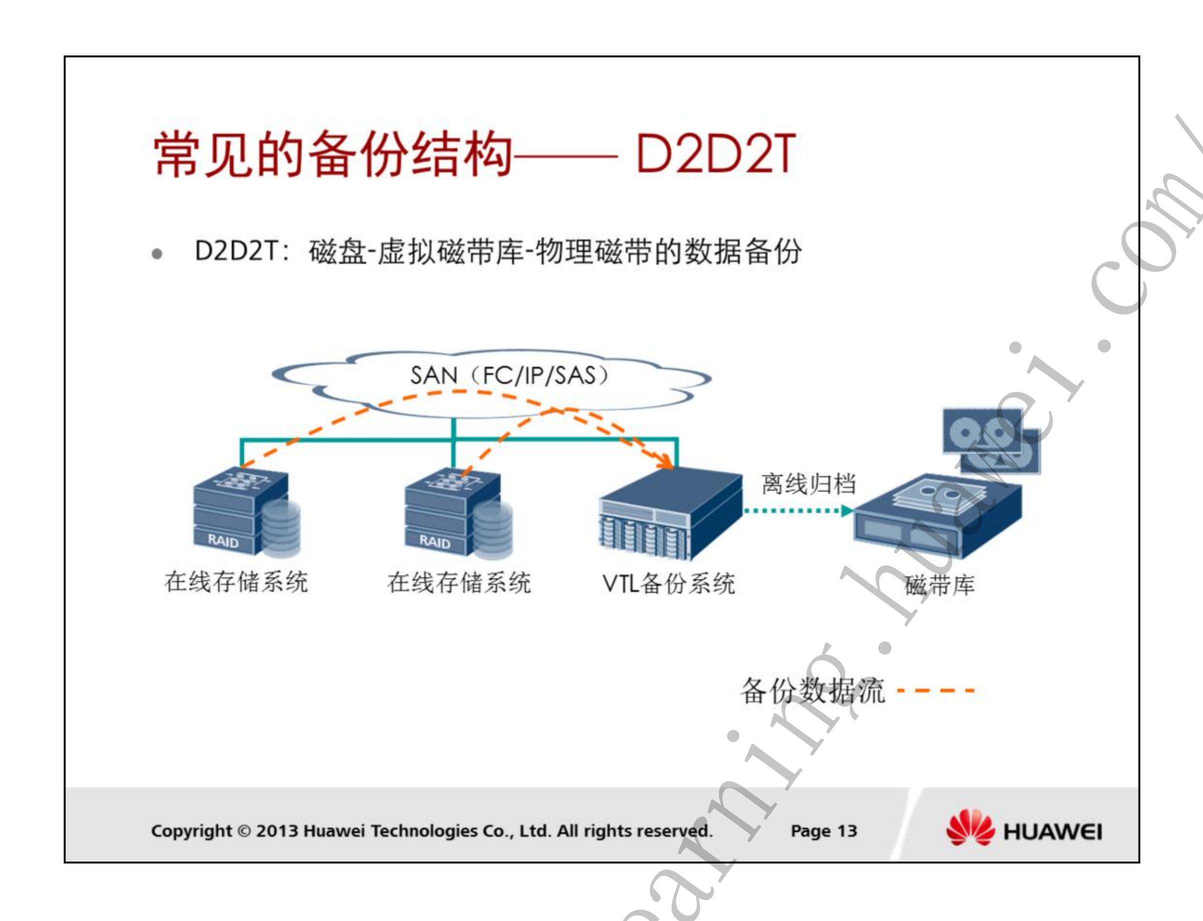
物理磁带的每盘容量都是固定的，用户在创建备份策略时，往往是确定某几盘磁带分别进行增量或差别备份，而另外一部分进行全备份，这就出现了用于进行增量或差别备份的磁带存储容量利用率非常低的问题，在一定程度上造成了用户投资的浪费。



在D2V的备份结构中，虚拟磁带库是采用物理硬盘为存储介质，通过虚拟化引擎来实现机械手臂、磁带驱动器以及磁带插槽的全新备份介质解决方案。由于没有机械零部件，虚拟磁带库的可靠性和可维护性比起物理磁带库大大提高，与磁盘阵列的物理可靠性、可维护性相一致。虚拟磁带库采用了物理硬盘作为存储介质，物理硬盘的随机读写、高速寻道定位在性能上远远高于物理磁带的顺序读写；因此，虚拟磁带库的I/O性能取决于虚拟磁带库的对外连接带宽，而非物理磁带库的磁带驱动器类型及数量的总和换算结果。

虚拟磁带库由于采用了虚拟化引擎，在与之连接的服务器端看到的仍然是个物理磁带库，而物理磁带库在操作系统中只有通过特定的备份软件才能够对其进行读写操作，因此虚拟磁带库虽然采用物理硬盘存储数据，但却避免了采用磁盘阵列作为备份介质时，数据被误删除或感染病毒等问题。

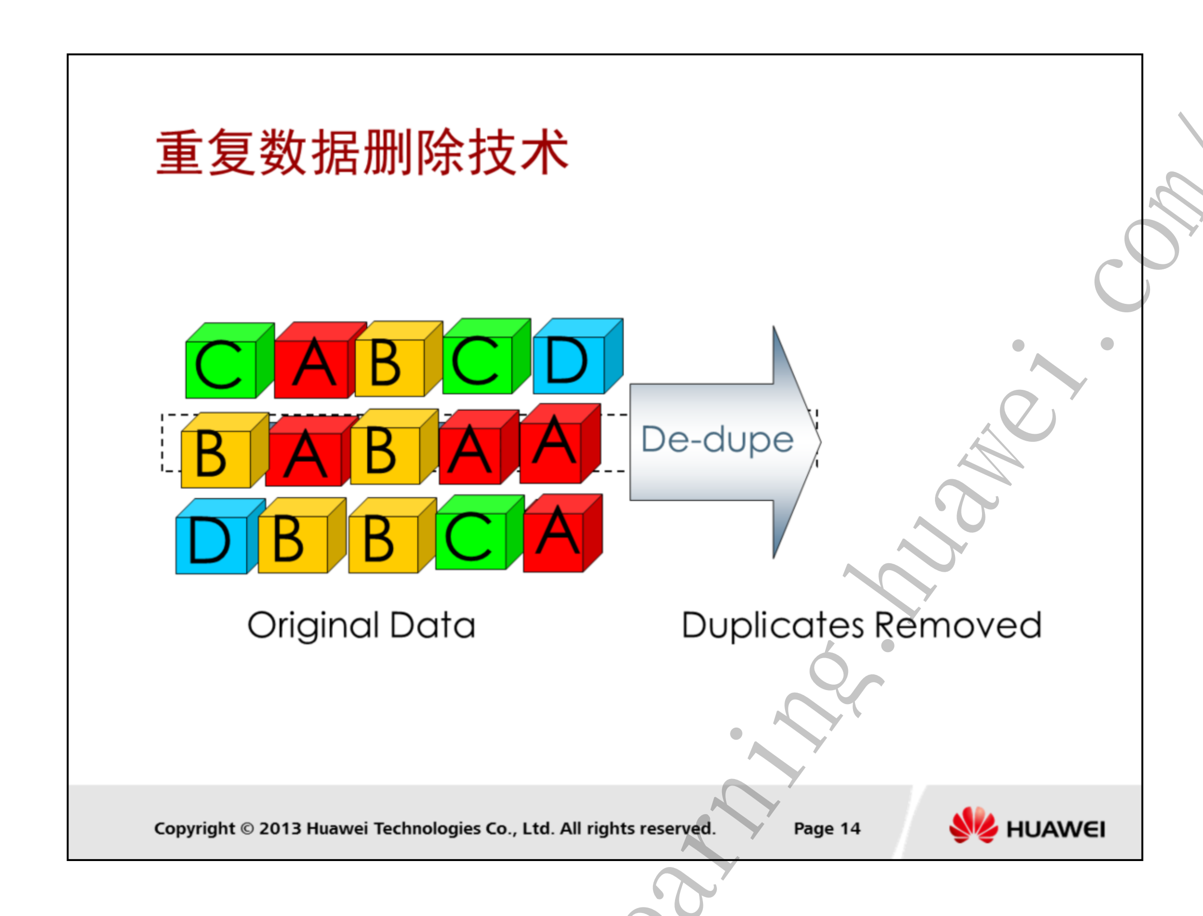
应用虚拟磁带库作为备份介质既保证了备份系统的可靠性又不增加任何投资的情况下提高了备份系统的备份效率；但是，用户还要考虑一些问题，首先，虚拟磁带库所保存的数据均是存在物理硬盘上的，而通常情况下，这些物理硬盘都是通过RAID技术组成的一个大的逻辑磁盘，如果用户需要将一部份重要的备份数据归档保存，这对于虚拟磁带库来说是个不小的挑战，因为用户无法去定位虚拟磁带具体存放在哪块物理硬盘中（在这物理磁带库中很容易实现，只需将某一盘对应的磁带从物理磁带库中取出即可）；其次、虚拟磁带库是否像物理磁带库一样具备数据压缩功能也是需要考察的；再次、虚拟磁带库是否能够提供按需存储的功能（即，虚拟磁带库可以控制某些用于增量或差别备份的虚拟磁带所占用的磁盘空间根据实际的存储数据量来设定）。



D2D2T的数据备份方式相对而言是最合适的备份方式，兼顾了可靠性、可管理性以及性能等多方面的因素。

虚拟磁带库有着安全、可靠、性能高的优点，而物理磁带库相对来说有着支持介质可移动的功能；综合分析来看，最好的解决方案应该是如下诸方面的整合：

- 采用物理磁盘作为一级备份介质，并通过RAID功能进行保护，以提高性能和可靠性。
- 采用虚拟化磁带库技术确保主机端备份系统的可管理性和安全性不受到挑战。
- 虚拟磁带库系统能够支持按需存储功能，实现存储资源的充分利用。
- 虚拟磁带库系统能够支持将虚拟磁带导出到物理磁带的功能，方便备份数据的归档保存及异地保存。



重复数据删除（Deduplication）技术，简单来说，就是一种消除重复数据的技术，它用软件或硬件的方式，对存储数据进行处理，以消除其中的重复数据，从而减小数据占用的存储空间。

- 重复数据删除的实现过程：
  - 原始数据存储在介质上。
  - 重删时以一定的数据块大小为单位进行比对。
  - 唯一的数据存放在重删后的空间里，后面的数据跟唯一数据进行比对，重复的数据则被删除，同时在指定的空间存放Index和Metadata。
  - 没有重复的数据则将这部分作为唯一数据保留在重删后的空间中，同时写入Index和Metadata。
- 对备份的价值：
  - 能更有效的节省存储的空间，大大提高存储的利用率，降低用户的TCO。
  - 减小备份窗口。

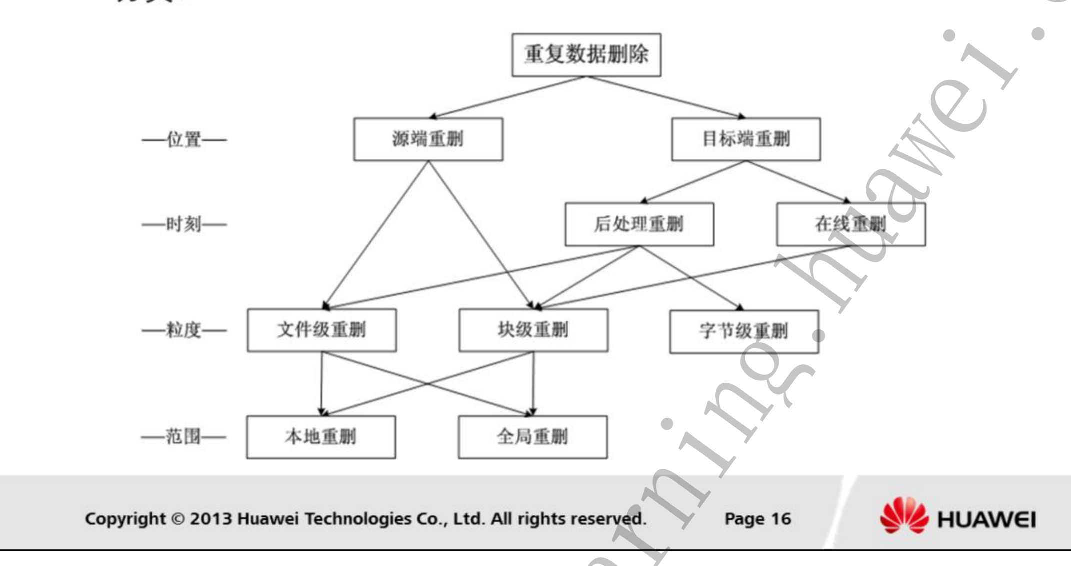
重复数据删除与压缩区别

比较项	功能	实现	数据内容	条件
重复数据删除	节省存储空间	切块比对，保留唯一数据	保留一份唯一数据	有基本的比对块
压缩		压缩算法	不改变原始数据内容	装压缩软件

 Page 15



- 可以把重复数据删除看作一种特殊的压缩。重复数据删除对数据进行一些算法的处理，把数据切割成块，比如说4K，或者是16K，32K，然后进行比对，比对完了以后只有变化过的，唯一的数据块才存到真正的磁盘空间上去，这是基本的原理。重复数据删除主要用于备份，进行重复数据删除，需要有一个基本的比较数据块。
- 压缩一般是通过压缩算法，减少文件的大小，删除文件的重复数据也许是文件压缩的一种方法，但压缩文件应该不仅仅限于此。



- 源端重删

  先删除重复数据，再将数据传到备份设备。

- 目标端重删

  先将数据传到备份设备，存储时再删除重复数据。

- 在线重删(Inline)

  数据写入磁盘之前进行重复数据删除。

- 后处理重删(Post-processing)

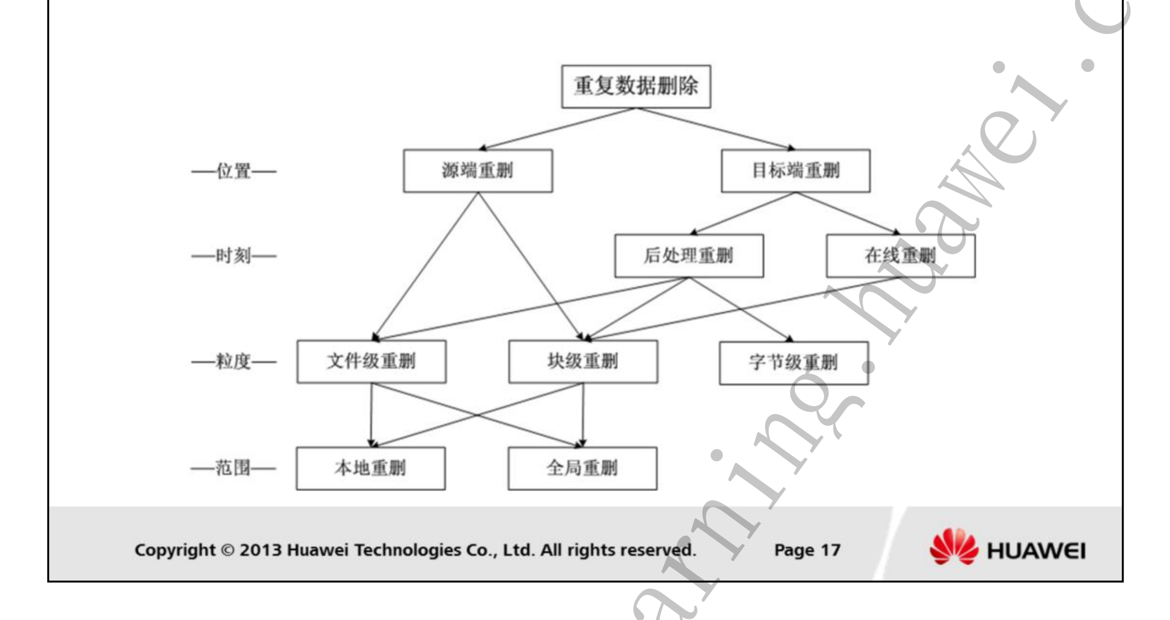
  数据写入磁盘后开始进行重复数据删除。

- 适应性重删(Adaptative Data Deduplication）

  性能要求较低的环境下采用在线重删，性能要求较高的时候采用后处理重删。

- 文件级重删

  也称为单实例存储(SIS)，根据索引检查需要存储文件的属性，并与已存储文件进行比较。如果没有相同文件，就将其存储，并更新索引；否则仅存入指向已存在文件的指针。



- 块级重删

将文件/对象分解成固定长度或不定长度的数据块，对数据块进行哈希计算，并与已存块的哈希值进行比较，只存储唯一哈希的数据块。

- 字节级重删

从字节层次查找和删除重复的内容，一般是通过压缩算法对用户数据进行压缩存储。

- 本地重删

查找重复数据时，仅和当前存储设备内的数据进行比较

- 全局重删

查找重复数据时，和整个重删域中的所有存储设备的数据进行比较。

重复数据删除关键指标

用户的问题	重复数据删除关键指标
能节省多少空间？TCO能降低多少？	重删率
需要多长时间重删完，是否会影响备份窗口？	重删性能
重删后数据是否可靠，是否随时可恢复？	数据可靠性
容灾场景下，需要多长时间才DR-Ready？	复制性能
一旦生产数据丢失，恢复需要多长时间？	恢复性能



目录



备份策略的内容

项目	内容
数据类型	文件、操作系统、数据库 裸设备备份、备份软件日志……
备份介质	磁盘、磁带 备份服务器……
备份类型	全量备份、增量备份、差量备份
数据保留时间	一周，一个月 一年……
备份周期	每天备份 每周备份……
备份窗口	备份时间范围

Page 20
HUAWEI

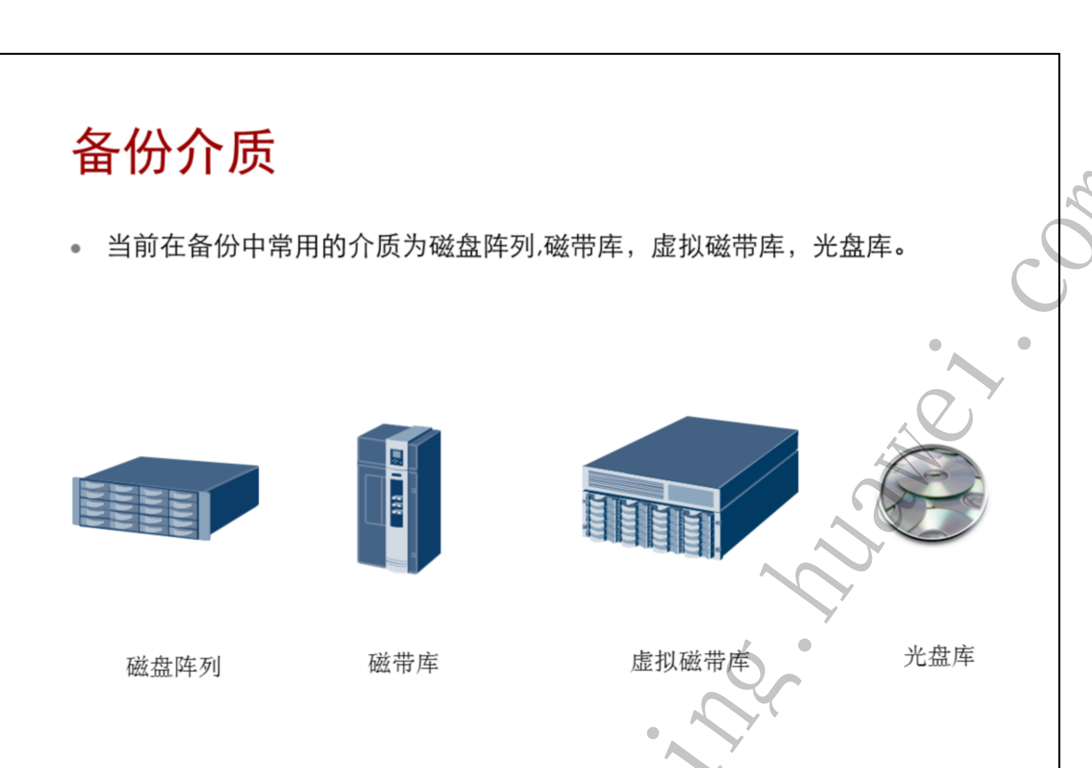
- 备份数据：即我们需要保护的数据。
- 备份目的地：即将保护的数据存放到指定的设备上，可以理解为备份介质。
- 备份类型：采用哪一种方式进行备份，分成全备份、增量备份、差异备份。
- 数据保留时间：存放在介质上的备份数据需要保留的时间，即备份数据有效时间。
- 备份周期：备份作业的频率，可以是每天、每周、每月等。
- 备份窗口：一次备份从开始到备份结束的时间段，称为备份窗口。
- 备份策略的选择：
  - 对于操作系统和应用软件，在每一次系统更新或者安装了新的软件后做一次全备份。
  - 对于关键的应用数据，涉及每天大量的数据更新，但是总数据量不是很大，我们可以用户在使用少量数据的时候每天做一个全备份。
  - 对于关键应用，且相对数据的总量每天只有少量的数据更新，我们可以在每月或每周做一个全备份，在此基础上可以在一系列短一点的间隔上做增量备份。

备份策略-数据类型

- 备份数据类型：文件、数据库、OS、应用软件…

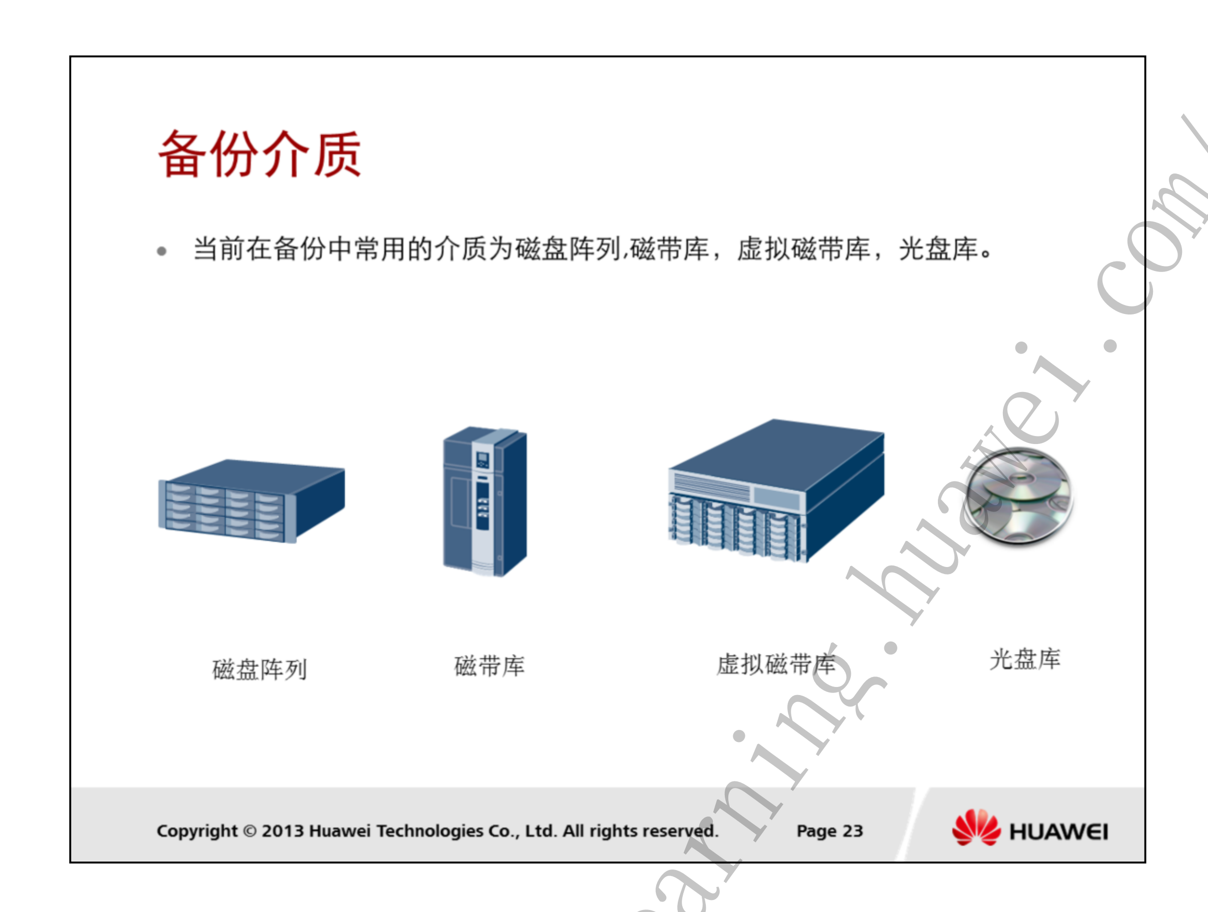
文件/文件夹备份	Word/excel/ppt/photo...
数据库备份	Oracle/db2/informix/sybase
逻辑卷备份	Oracle逻辑卷/MySQL逻辑卷
操作系统备份	Windows/redhat/suse···
备份软件自身的备份	Backup Exec/NetBackup···



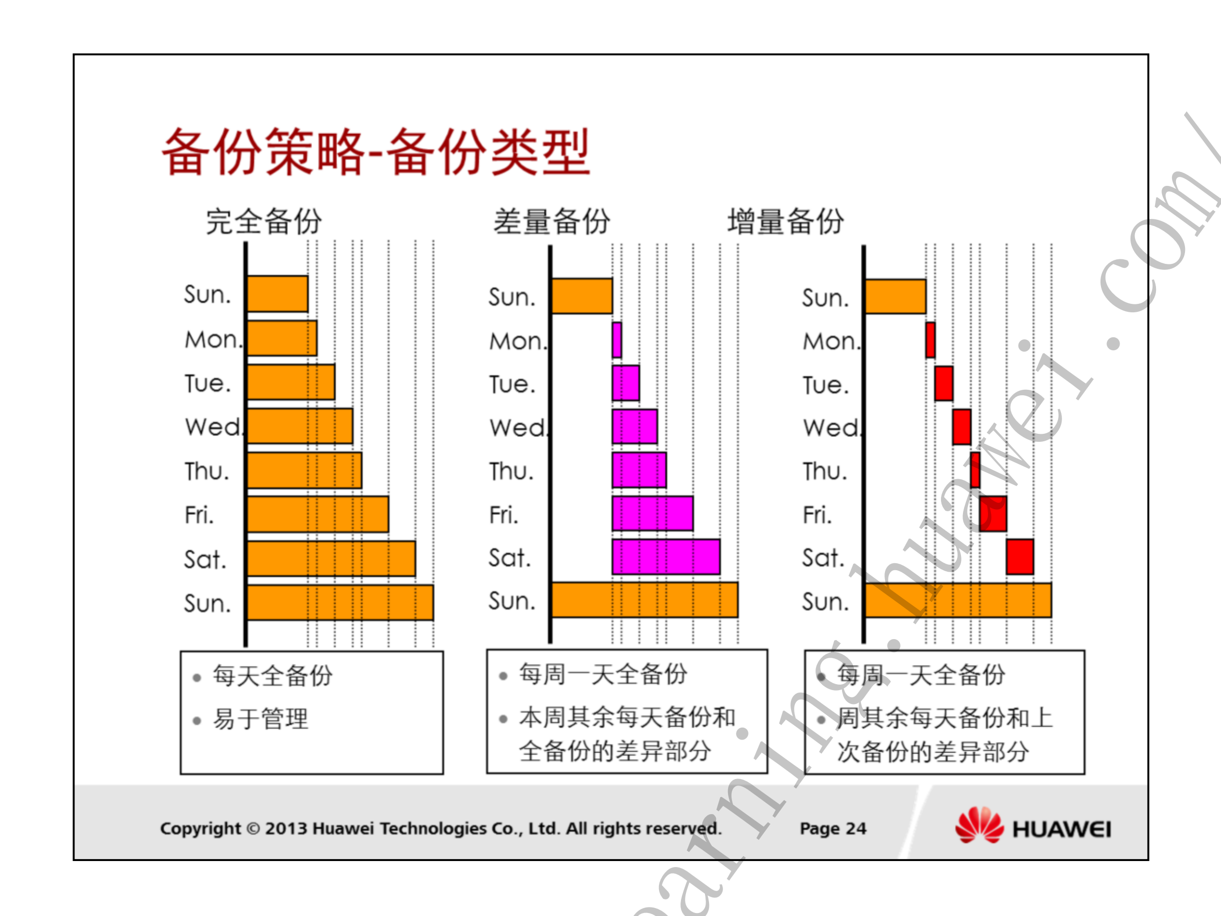


- 磁盘阵列做为备份介质
  - 优点：性能高，读写速度快。维护简单，可维护性较低。设备冗余较好（包括电源、风扇、磁盘阵列控制器均可冗余）。磁盘介质受外界环境（温度、湿度、粉尘等）影响相对较小。磁盘阵列可采用RAID保护。
  - 缺点：一次性投入成本较高。存储效率有待提高。磁盘备份无法防止人为错误操作。
- 物理磁带库
  - 优点：基于磁带的存储系统；由多个驱动器、槽、机械手臂和磁带构成；单位存储成本低；可实现数据和读写装置分离，理论上存储容量无限大。
  - 缺点：

    纯机械装置，硬件机械故障率较高；磁带介质本身较脆弱，容易受到外界环境（温度、湿度、粉尘等）影响而失效；管理、维护成本较高；设备冗余性较差（通常只有大型磁带库有电源冗余）；备份/恢复时间较长；只能顺序读写，不能随机读写；
- VTL：Virtual Tape library虚拟磁带库
  - 优点：管理方便；性能较高；无需更改原有磁带存储系统架构；存储性能较高，可采取压缩、重复数据删除等技术。
  - 缺点：存储介质为磁盘导致单位存储成本较高；整体部署成本较高；容量扩展性能较真实磁带库差。



- 光盘库/光盘塔
  - 优点：光驱、光盘的价格较低，具备成本优势；光盘介质保存时间长；对保存环境要求较低。
  - 缺点：读/写速度慢；光驱数量有限，数据源少，所支持的用户数量也较少；通常不可反复擦写。



- 完全备份：
  - 优点：能够基于上一次的全备份快速恢复数据，恢复窗口小。
  - 缺点：所占用的存储空间大，每次备份耗时长，备份窗口大。
- 差量备份：
  - 优点：相对全备份来说每次备份可以节约一个全备份的存储空间，备份窗口较小，恢复窗口较小。
  - 缺点：恢复时必须依赖上一次全备份和本次的差量备份才能完整恢复数据。
- 增量备份：
  - 优点：能够最大限度地节省存储空间，备份窗口小。
  - 缺点：数据恢复时必须依赖上一次完全备份和每一次的增量备份才能对数据进行完整恢复，恢复时数据重构较慢，恢复窗口较大。

备份策略-保留周期

- 保留周期Retention 即在介质上存放的备份数据的有效期，在保留周期内的数据是不允许被覆盖，当数据存放时间超过保留周期后，该部分数据所使用的介质空间可以被覆盖，从而释放介质空间。

Dispose
Create
Protect
Access
Migrate
Archive
数据生命周期

Page 25
HUAWEI

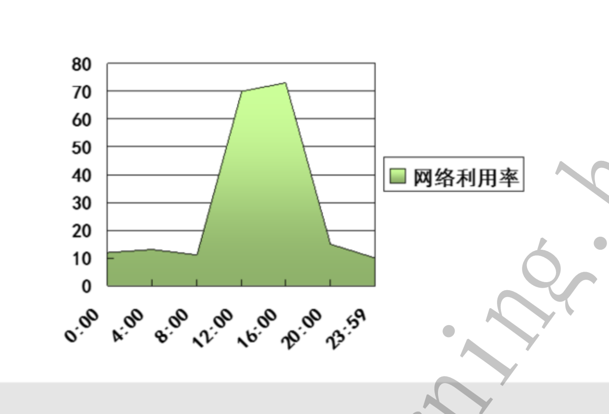
系统管理员可以指定每次备份可以保留多长时间，当该期限到达时，备份软件自动将该备份的相关信息从备份软件数据库中删除（并不从磁带、磁盘中删除），这时用户就检索不到这次备份的信息。

数据被“创建”生成后，重要的数据需要“保护”起来，且重要的数据需要经常的“访问”，当此部分数据的重要性降低后，需要将这部分数据“迁移” 到一些大容量、性能较低的存储设备上，随着时间的推移，当该部分数据的重要性再次降低，此时需要将这部分数据“归档”，在一定的期限（保留周期）后，这部分数据可以被“去除”，即使备份集失效。

备注：备份集（Backup sets）顾名思义就是一次备份的集合，它包含本次备份的所有备份片。一个备份集根据备份的类型不同，可能构成一个完全备份或增量备份。

备份策略-备份窗口

- 备份窗口（Backup window ）是指在不严重影响使用需要备份的数据的应用程序情况下，进行数据备份的时间间隔，也就是完成一次给定备份所需的时间。





 Page 26 HUAWEI

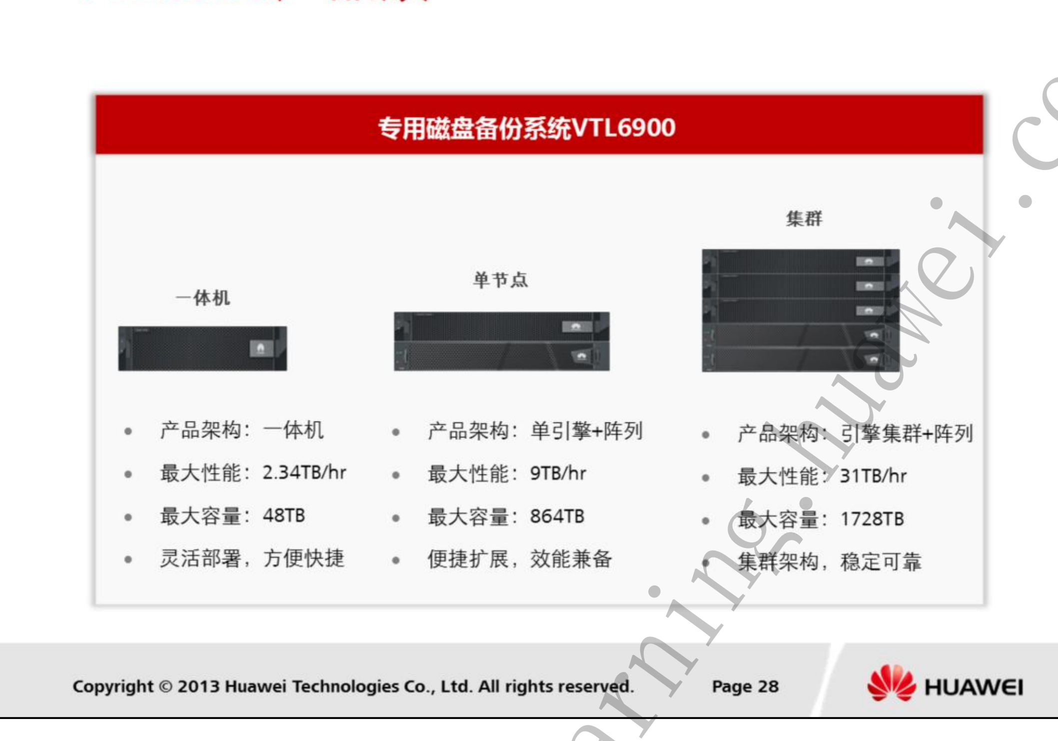
业务连续性（BC）要求越高，备份窗口与之矛盾也就越突出，备份系统需要在业务连续性和备份窗口之间作出相应的平衡。

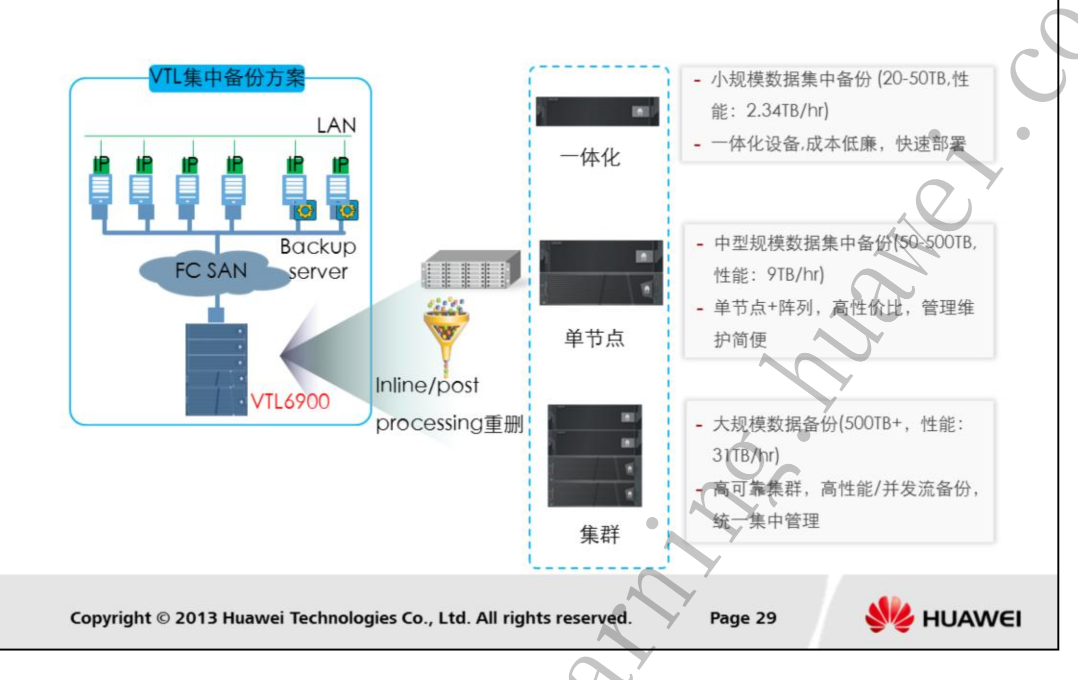
图示的网络利用率表示：在8:00—20:00网络利用率是最大的时候，此时间段不适合做备份窗口，因为此时备份会对用户的业务运行造成影响，所以在选择和设置备份窗口时，需要设置在网络利用率不高的时间段。



目录



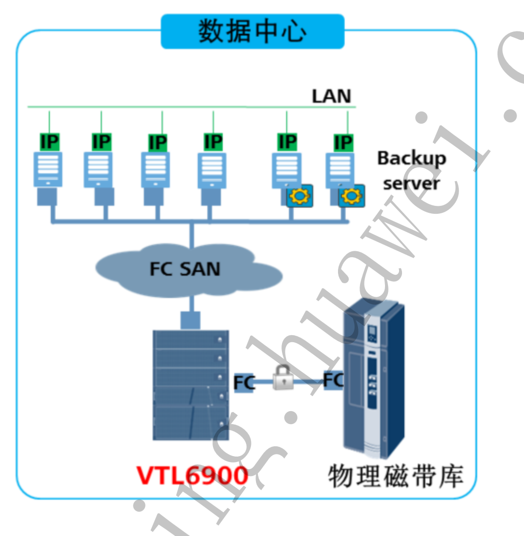

VTL6900产品族
专用磁盘备份系统VTL6900
一体机
单节点
集群
产品架构：一体机
最大性能：2.34TB/hr
最大容量：48TB
灵活部署，方便快捷
产品架构：单引擎+阵列
最大性能：9TB/hr
最大容量：864TB
便捷扩展，效能兼备
产品架构：引擎集群+阵列
最大性能：31TB/hr
最大容量：1728TB
集群架构，稳定可靠
Copyright © 2013 Huawei Technologies Co., Ltd. All rights reserved.
Page 28
HUAWEI



- 中小局点
    - 容量：20TB ~160TB
    - 保存：1~6个月
    - 性能：400MB/s~1250MB/s
    - 预算：比较有限

VTL备份归档一体化方案

- 场景需求
  - 大量历史数据需要长期保存 （保存：6个月+）
  - 物理磁带库备份性能低下
  - 备份管理维护繁琐
  - 预算有限，设备利旧，保护已有投资
- 客户价值
  - 缩短备份窗口，VTL6900作为高性能归档缓存
  - 设备利旧，已有物理带库作为大容量归档存储
  - 简化管理，VTL6900自动将备份数据归档至磁带库



 Page 30 

- 分级备份
  - 利用已有物理带库
  - 原备份性能不高<200MB/s
  - 备份数据长期保存--超过12个月

HDP3500E 简介

- 设备简介
  - HDP3500E 是一款集备份软件、备份服务器和备份介质于一体的高性能备份产品。
  - 集成NetBackup 备份软件，能够为企业关键业务提供全面的数据保护。

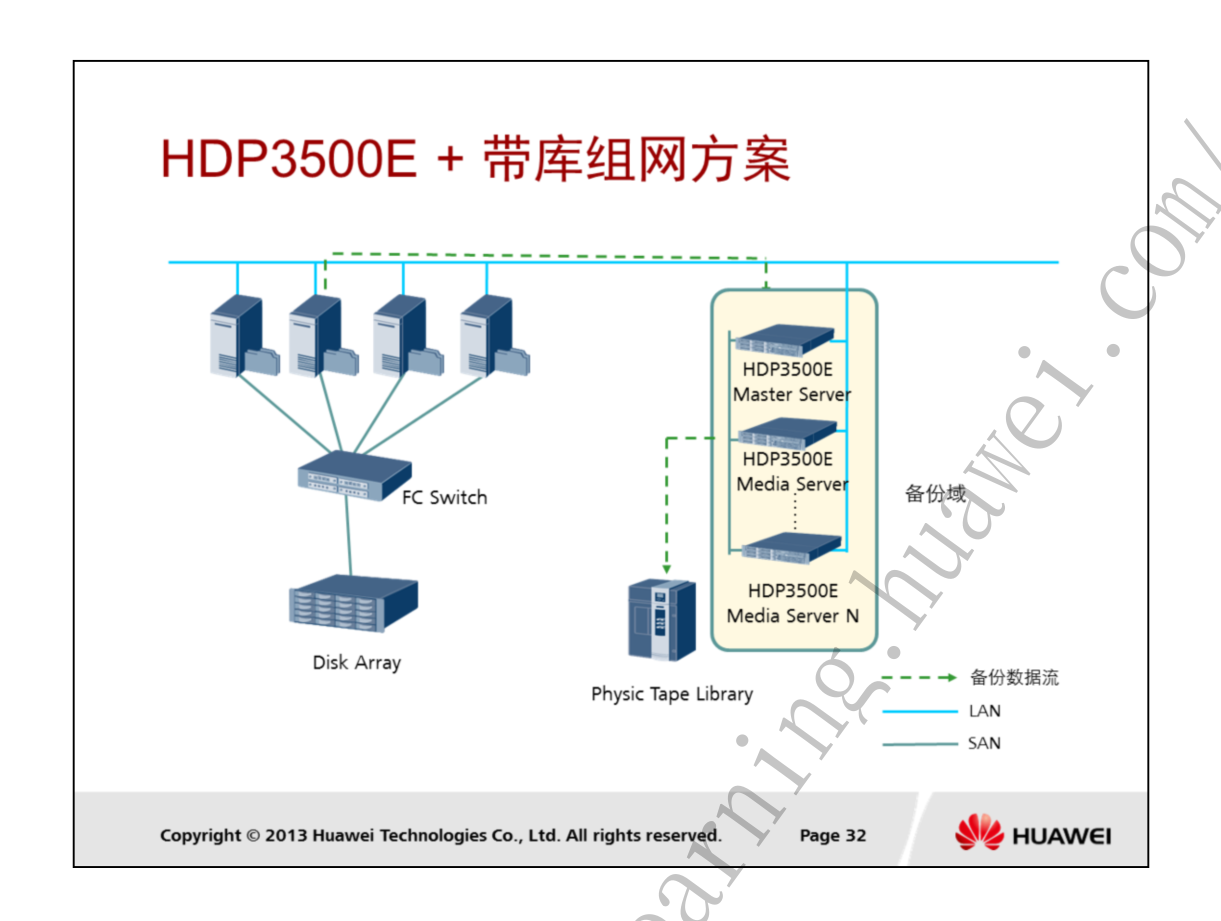


 Page 31



2U，12盘位。可提供18TB备份可用容量，以及4个千兆备份业务网口。

通过备份设备的横向扩展，用户可组建基于 HDP3500E 的备份域，实现备份容量和备份性能的线性增长。



该方案由多台HDP3500E组成一个备份域，1台Master Server，多台Media Server。

备份数据传输采用LAN-Based的方式，通过LAN网络承载。备份数据首先存储在HDP3500E的本地硬盘存储上，再根据用户需求，定期迁移到物理带库上，实现备份数据的分级存储，提高存储利用率、降低整体成本。

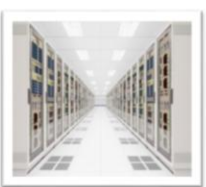
当业务增长后，可以通过增加HDP3500E设备，实现扩展备份性能、备份数据存储空间平滑扩展。通过增加HDP3500E，实现备份系统计算性能、存储空间的平滑扩展。可外接物理带库，实现备份数据分级存储，提高存储利用率。外接带库时，支持Vault功能，满足磁带离线管理需求。



目录



容灾建设背景



数据集中，风险加剧

- 政务信息化建设高速发展
- 政府行业逐渐建立大型数据中心，数据的集中也意味着风险的加剧
- 提高数据的抗风险能力，已成为急需解决的问题

业务中断，数据丢失，影响巨大

- 关键业务中断对政务信息化影响巨大
- 数据丢失在社保、财政、政务、高法等系统中导致不可弥补的损失。

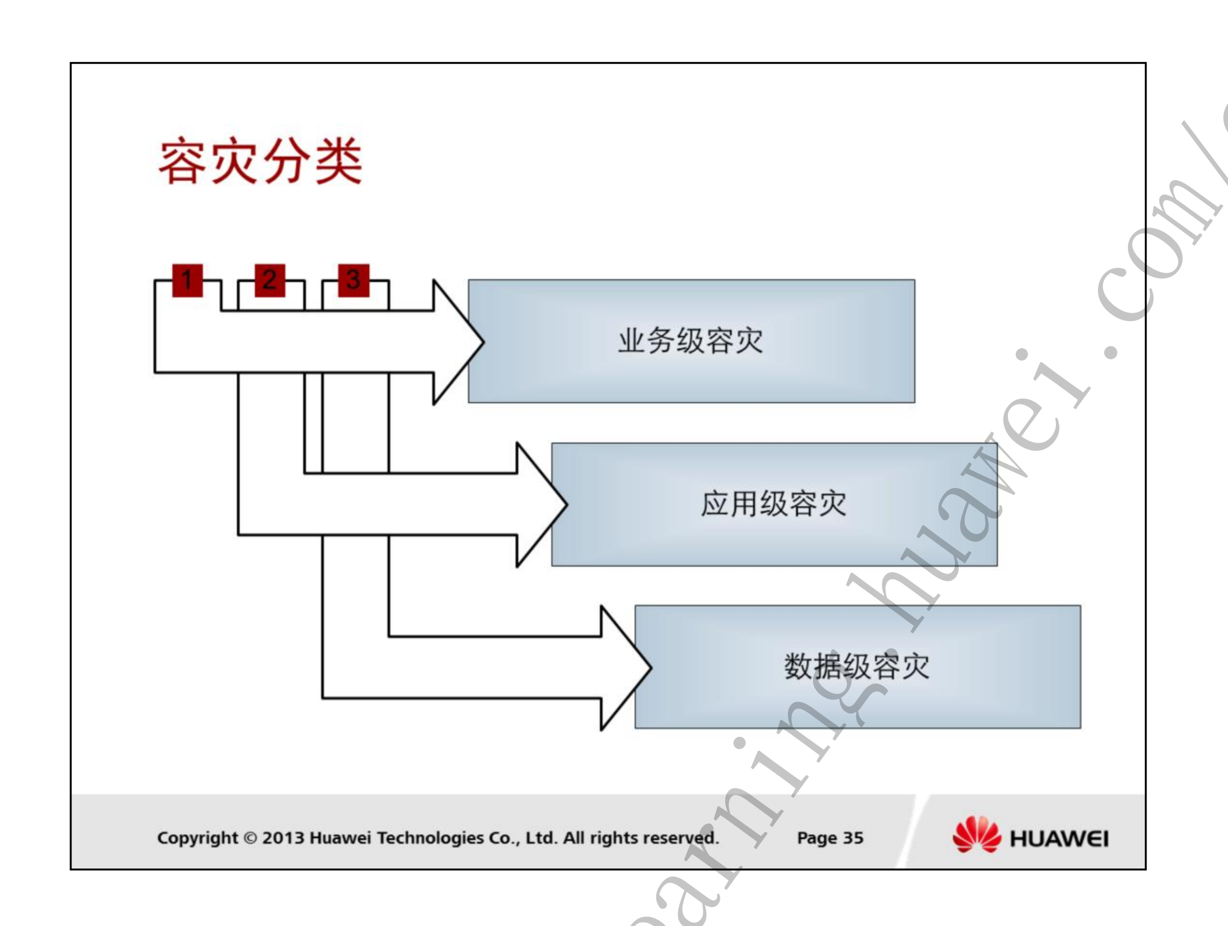


政策所导，形势所需

- 《国家信息化领导小组关于加强信息安全保障工作的意见》
- 《关于做好国家重要信息系统灾难备份的通知》
- 国家标准《信息系统灾难恢复规范》



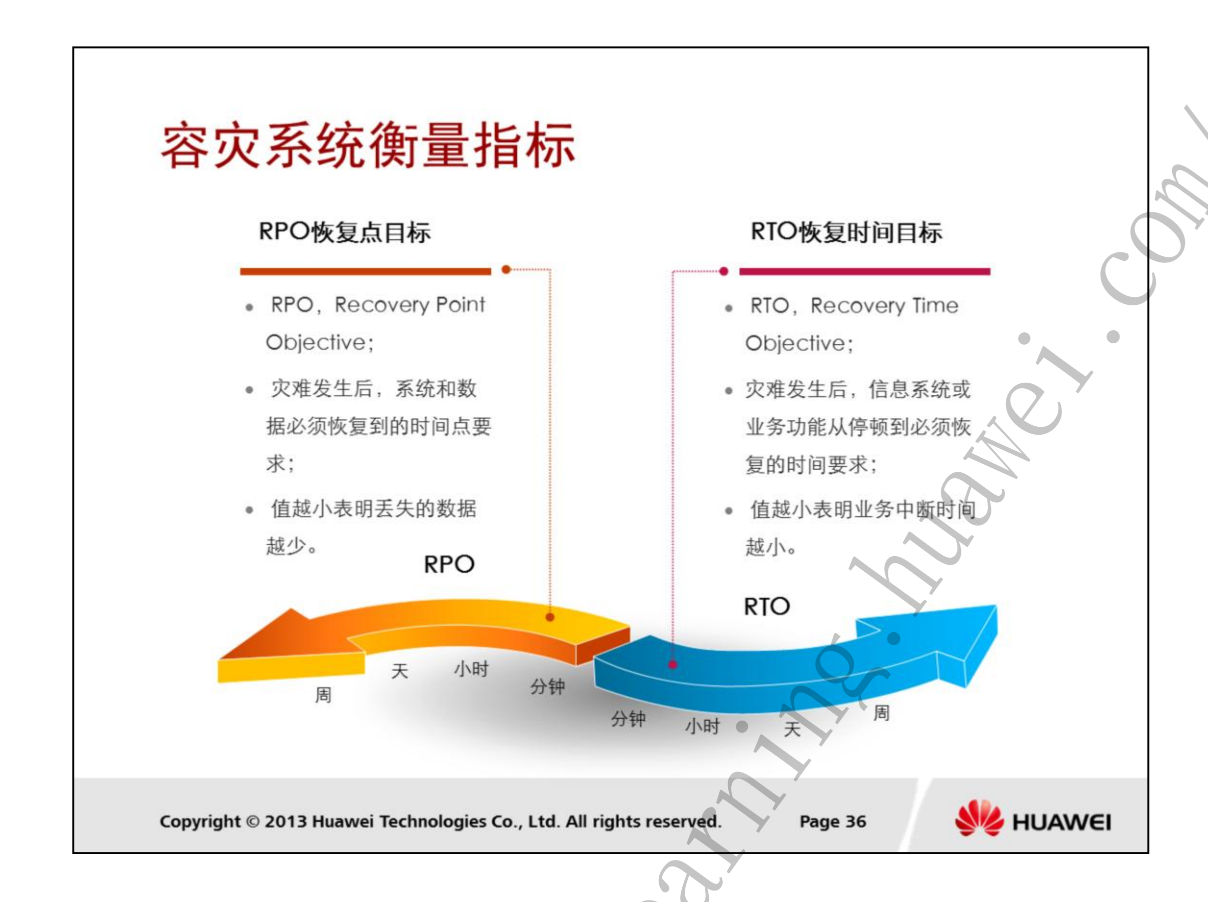




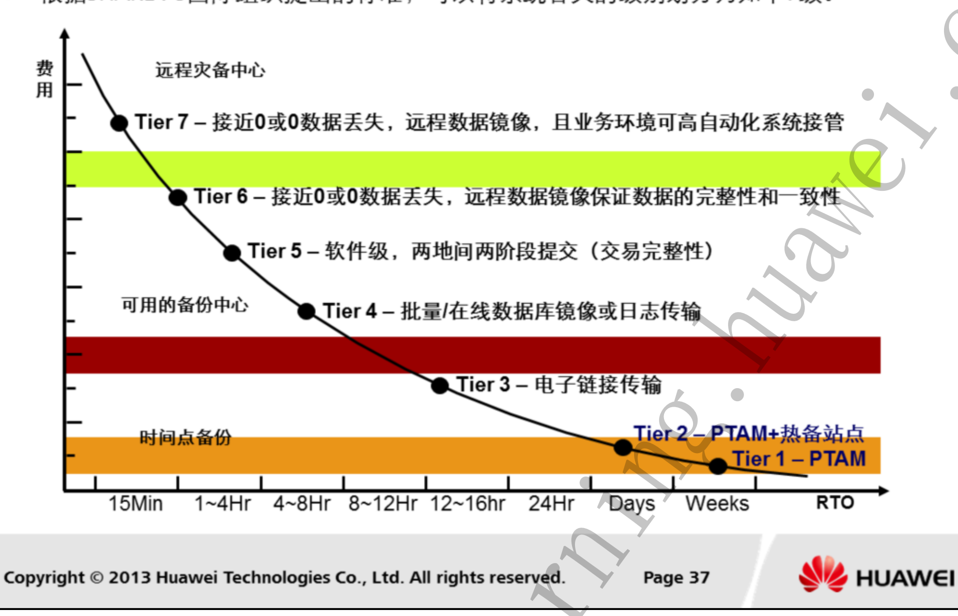
数据级容灾是指通过建立异地容灾中心，做数据的远程备份，在灾难发生之后要确保原有的数据不会丢失或者遭到破坏，但在数据级容灾这个级别，发生灾难时应用是会中断的。在数据级容灾方式下，所建立的异地容灾中心可以简单地把它理解成一个远程的数据备份中心。数据级容灾的恢复时间比较长，但是相比其他容灾级别来讲它的费用比较低，而且构建实施也相对简单。

应用级容灾是在数据级容灾的基础之上，在备份站点同样构建一套相同的应用系统，通过同步或异步复制技术，这样可以保证关键应用在允许的时间范围内恢复运行，尽可能减少灾难带来的损失，让用户基本感受不到灾难的发生，这样就使系统所提供的服务是完整的、可靠的和安全的。应用级容灾生产中心和异地灾备中心之间的数据传输是采用异类的广域网传输方式；同时应用级容灾系统需要通过更多的软件来实现，可以使多种应用在灾难发生时可以进行快速切换，确保业务的连续性。

业务级容灾是全业务的灾备，除了必要的IT相关技术，还要求具备全部的基础设施。其大部分内容是非IT系统（如电话、办公地点等），当大灾难发生后，原有的办公场所都会受到破坏，除了数据和应用的恢复，更需要一个备份的工作场所能够正常的开展业务。



- RPO用来衡量容灾系统的业务恢复能力。
- RTO用来衡量容灾系统的数据冗余备份能力。



PTAM（Pickup Truck Access Method）卡车运送访问方式

1层：有数据备份，无备用系统（Data Backup with No Hot Site）

2层：有数据备份，有备用系统 （Data Backup with Hot Site）

3层：电子链接（Electronic Vaulting）

使用3层容灾解决方案的业务，是在2层解决方案的基础上，又使用了对关键数据的电子链接技术。

4） 应用程序层：应用程序开发时考虑到数据的复制。 RTO小时级，快照等

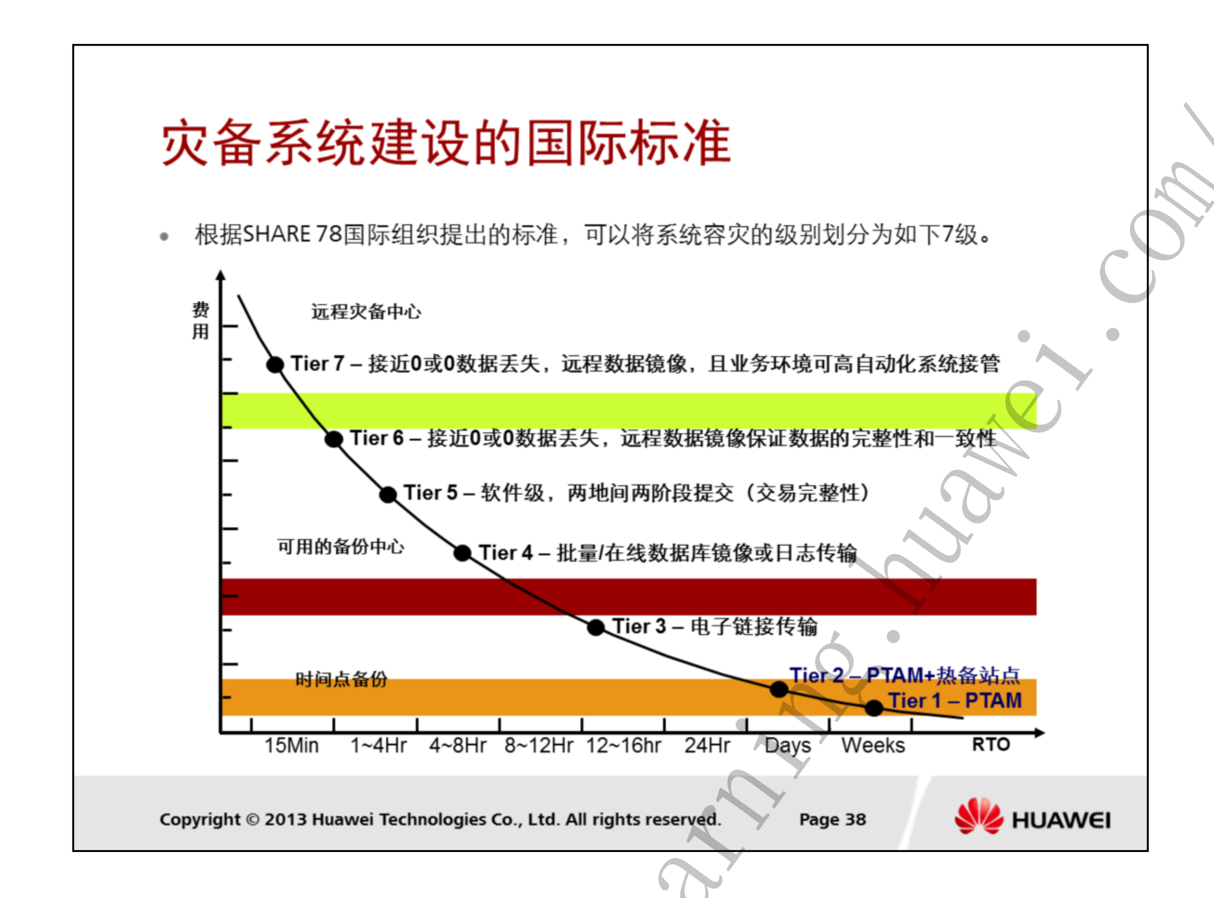
在4层方案中开始使用基于磁盘的解决方案。此时仍然会出现几个小时的数据丢失，但同基于磁带的解决方案相比，通过加快备份频率，使用最近时间点的快照拷贝恢复数据会更快。 系统可在一天内恢复。

5 ）层：交易的完整性 （Transaction Integrity）

5层除了使用4层的技术外，还要维护数据的状态 - 要保证在本地和远端数据库中都要更新数据。 只有当两地的数据都更新完成后，才认为此次交易成功。 生产中心和备用中心是由高速的宽带连接的，关键数据和应用同时运行在两个地点。

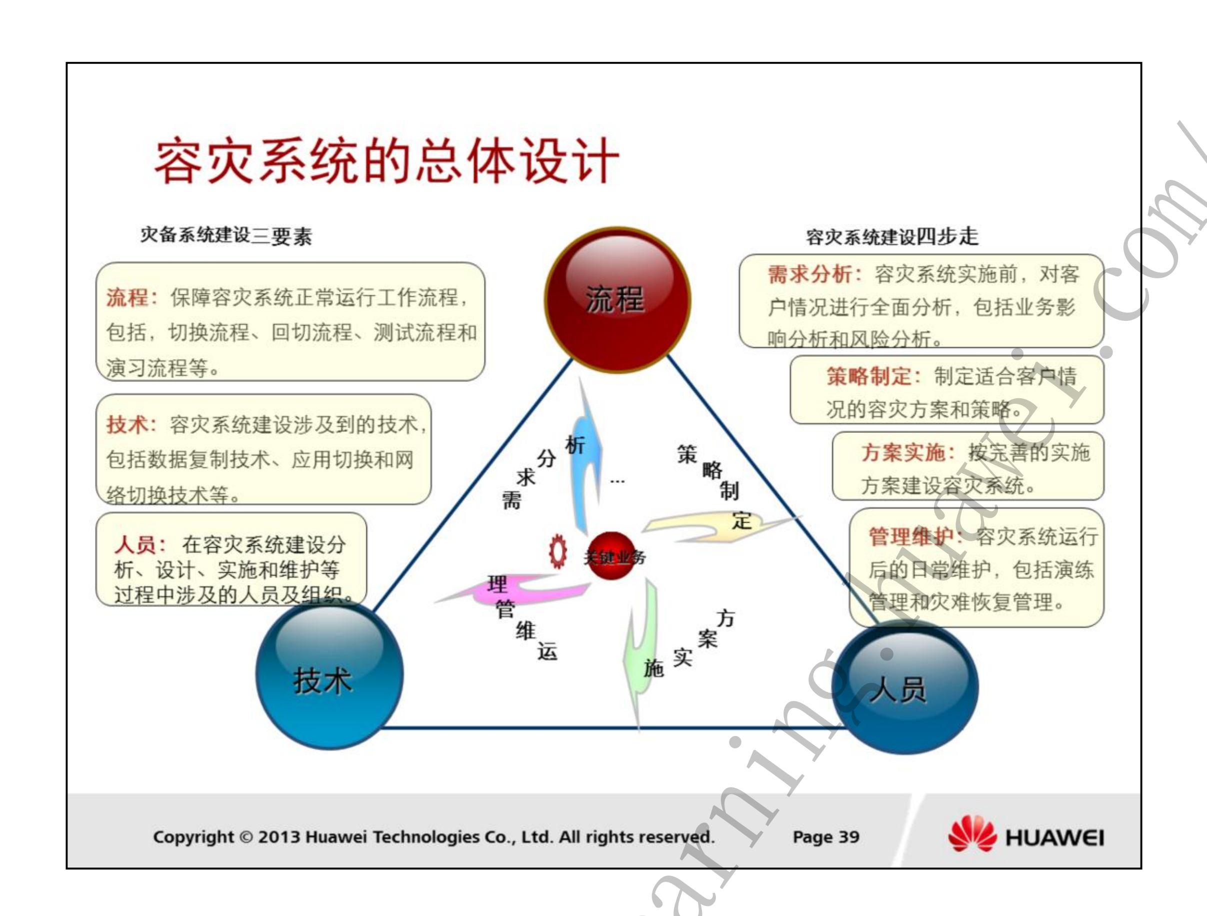
6）层：少量或无数据丢失 （Zero or little data loss）

6层灾难恢复方案可以保证最高一级数据的实时性。 适用于那些几乎不允许数据丢失并要求能快速将数据恢复到应用中的业务。 此种解决方案提供数据的一致性，不依赖于应用而是靠大量的硬件技术和操作系统软件来实现的。



7）层：解决方案与具体业务相结合，实现自主管理（Highly Automated，Business Integrated Solution)无需人工介入的自动站点故障切换功能

7层灾难恢复方案在第6层的基础上，集成了自主管理的功能。在保证数据一致性的同时，又增加了应用的自动恢复能力，使得系统和应用恢复的速度更快、更可靠（按照灾难恢复流程，手工操作也可实现整个恢复过程）。



流程、技术、人员是容灾系统建设的三个要素，贯穿容灾系统建设的全过程。

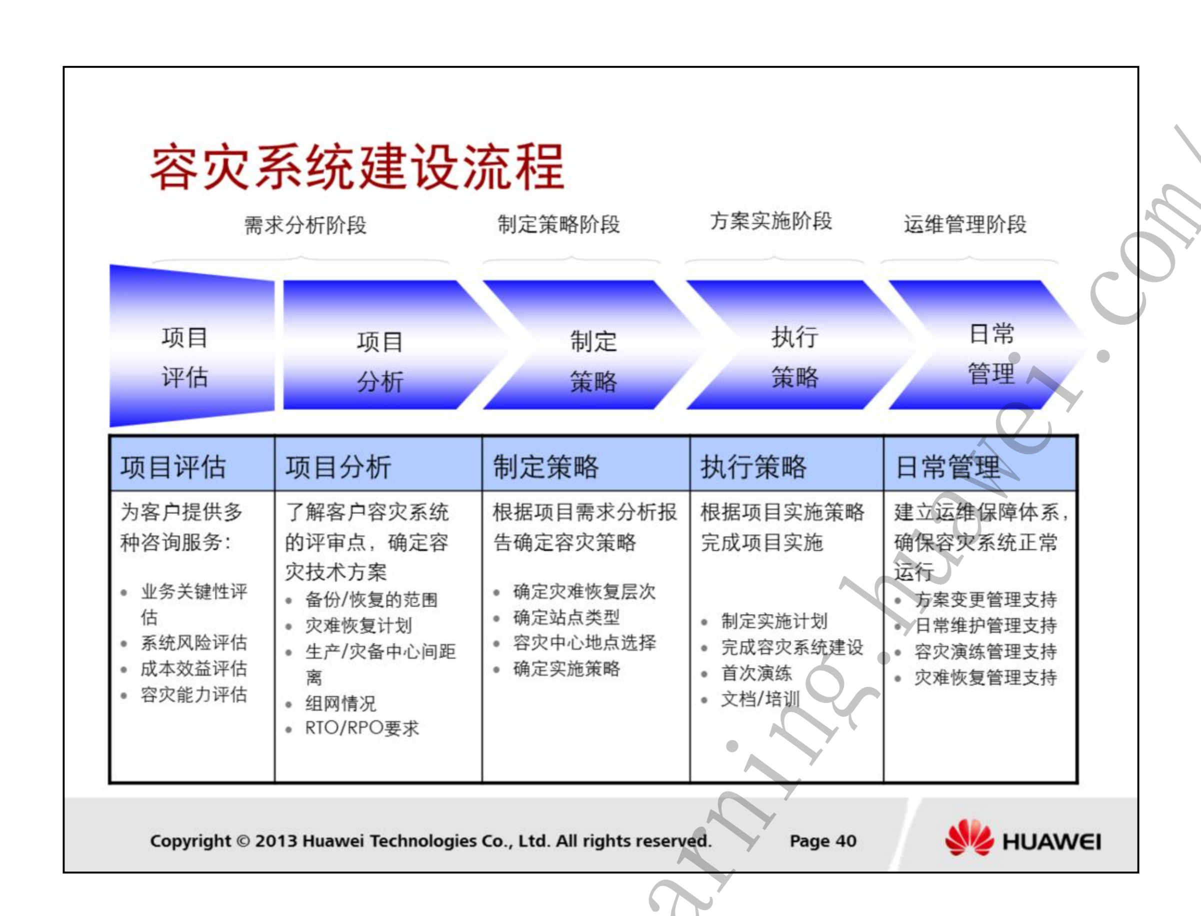
容灾系统建设分为四个阶段：

第一阶段 市场评估/需求分析

第二阶段 策略制定

第三阶段 方案实施

第一阶段 管理维护



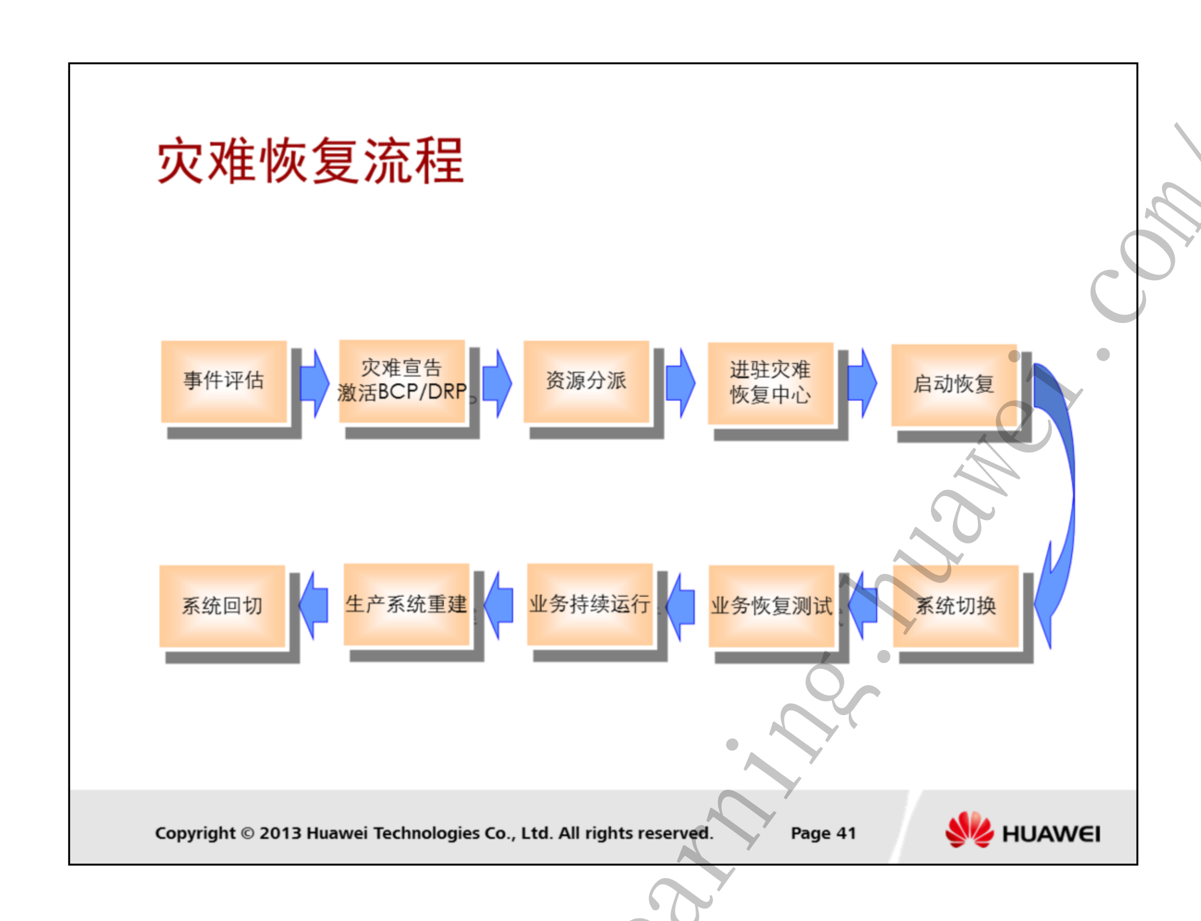
华为公司在容灾系统建设中实现全流程管理，在项目需求分析阶段，可提供项目咨询服务，在项目投入运营以后，继续提供技术支持服务。

方案变更管理：已有方案优化、扩容升级、新增业务

日常维护管理：硬件设备、资源监控、用户管理

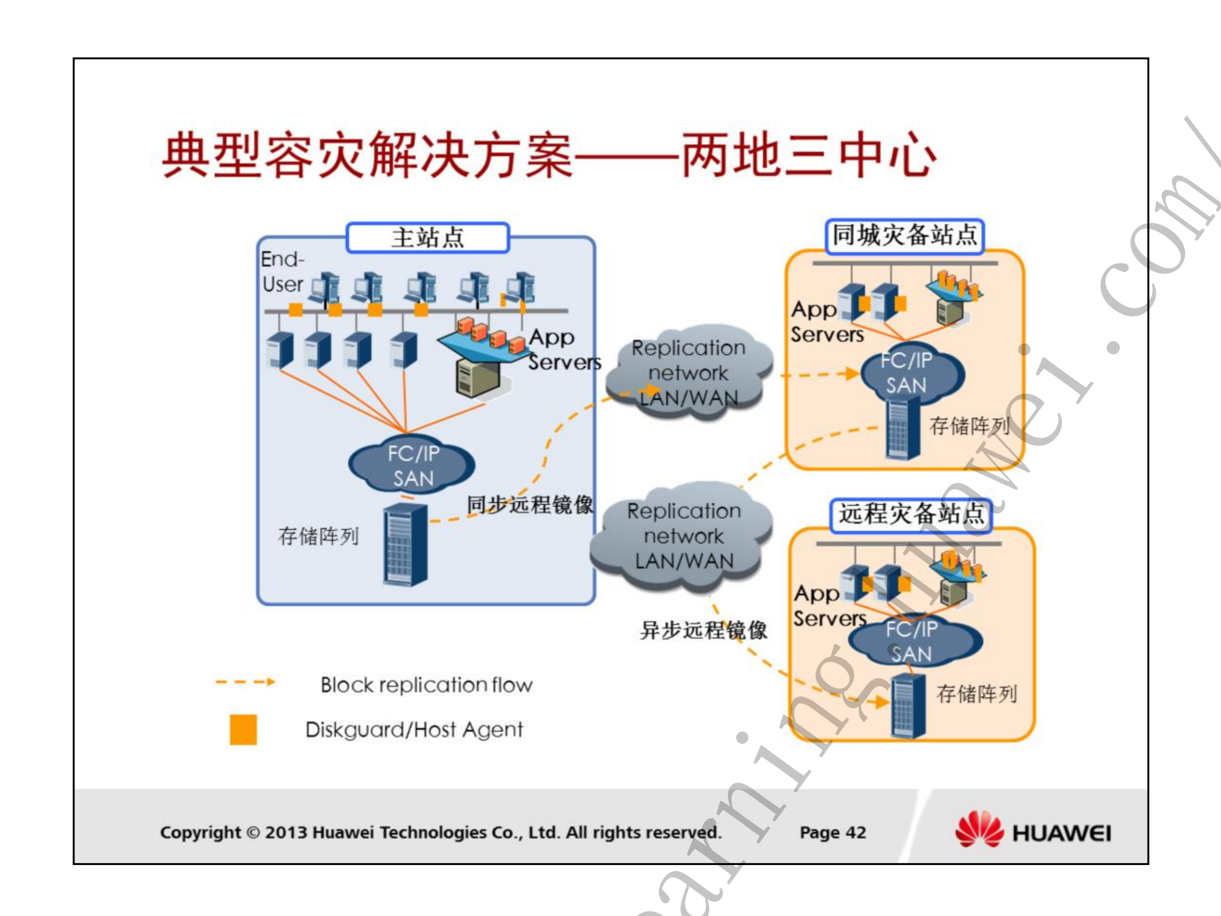
容灾演练管理：变更演练、定期演练、应急演练

灾难恢复管理：灾难监控、灾难恢复、容灾切换



BCP: Business Continuity Plan； 中文：业务持续性计划BCP是组织为避免关键业务功能中断，减少业务风险而建立的一个控制过程，它包括对支持关键功能的人力、物力和关键功能所需的最小级别服务水平的连续性保证。

DRP灾难恢复计划（Disaster Recovery Planning） 灾难恢复计划是一个全面的状态，它包括在事前，事中，和灾难对信息系统资源造成重大损失后所采取的行动。灾难恢复计划是对于紧急事件的应对过程。在中断的情况下提供后备的操作，在事后处理恢复和抢救工作



- 业务可靠性要求非常高，主站点故障后业务可以自动切换到同城灾备站点。同时，数据可以在异地保存，进行数据级恢复。通过磁盘阵列的级联复制功能，将数据复制到同城的容灾站点和异地的灾备站点。同城站点采用同步复制，保证数据的实时同步。异地站点采用异步复制，进行数据级容灾。



总结

- 备份概念及拓扑结构
- 备份技术
- 备份策略制定
- 华为备份实现与应用
- 容灾简介





1. 备份的拓扑结构有哪些，优缺点分别是什么？
2. 重复删除技术的分类?





练习题

- 多选题

1、常用的备份介质有（　　　　）

A. 磁带库

B. 磁盘阵列

C. 虚拟磁带库

D. 光盘塔/光盘库

2、重删技术可以按照重删的粒度可分为（　　　　）

A. 文件级重删

B. 块级重删

C. 字节级重删

D. 源端重删



- 习题答案

多先题

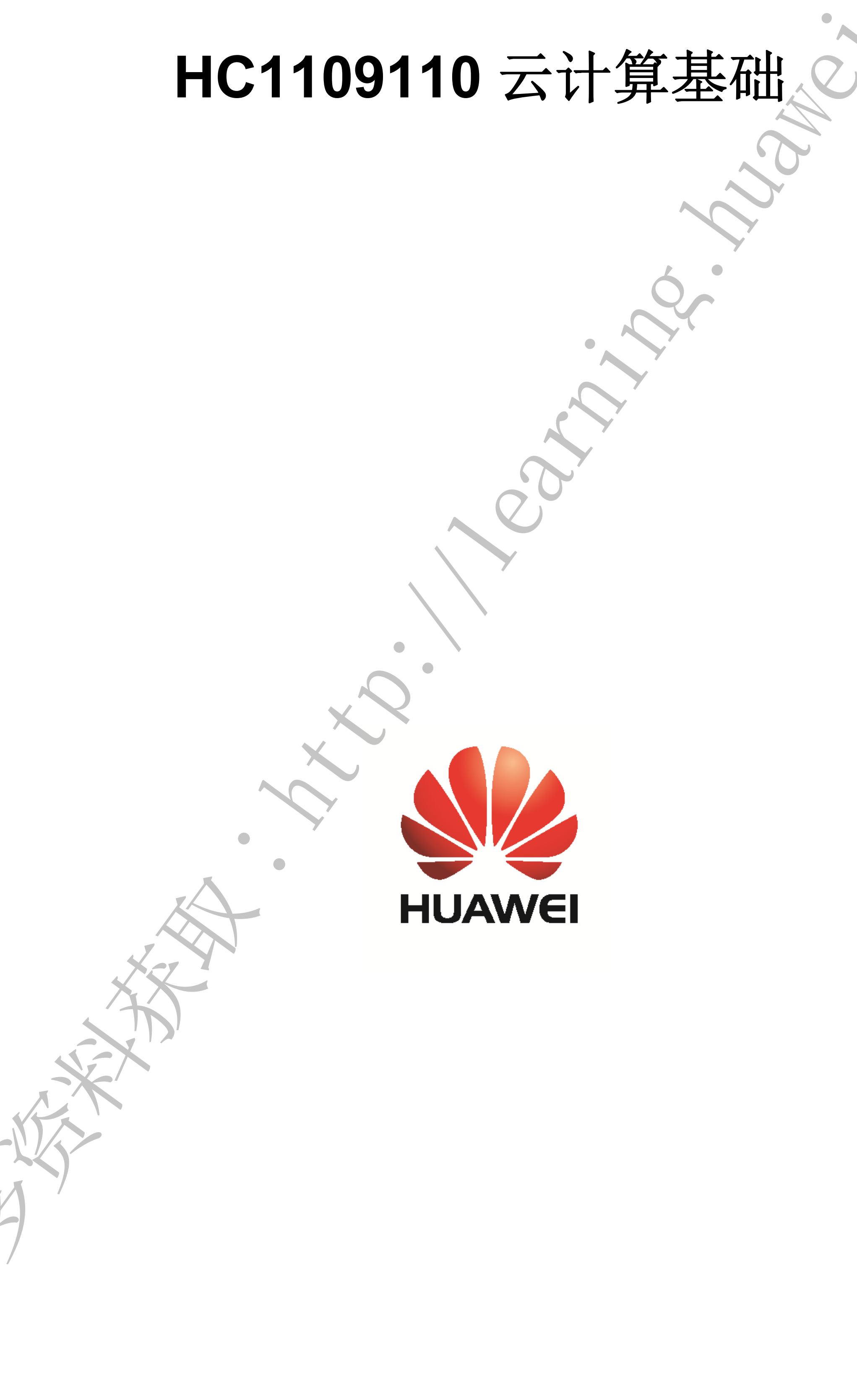
1、A B C D

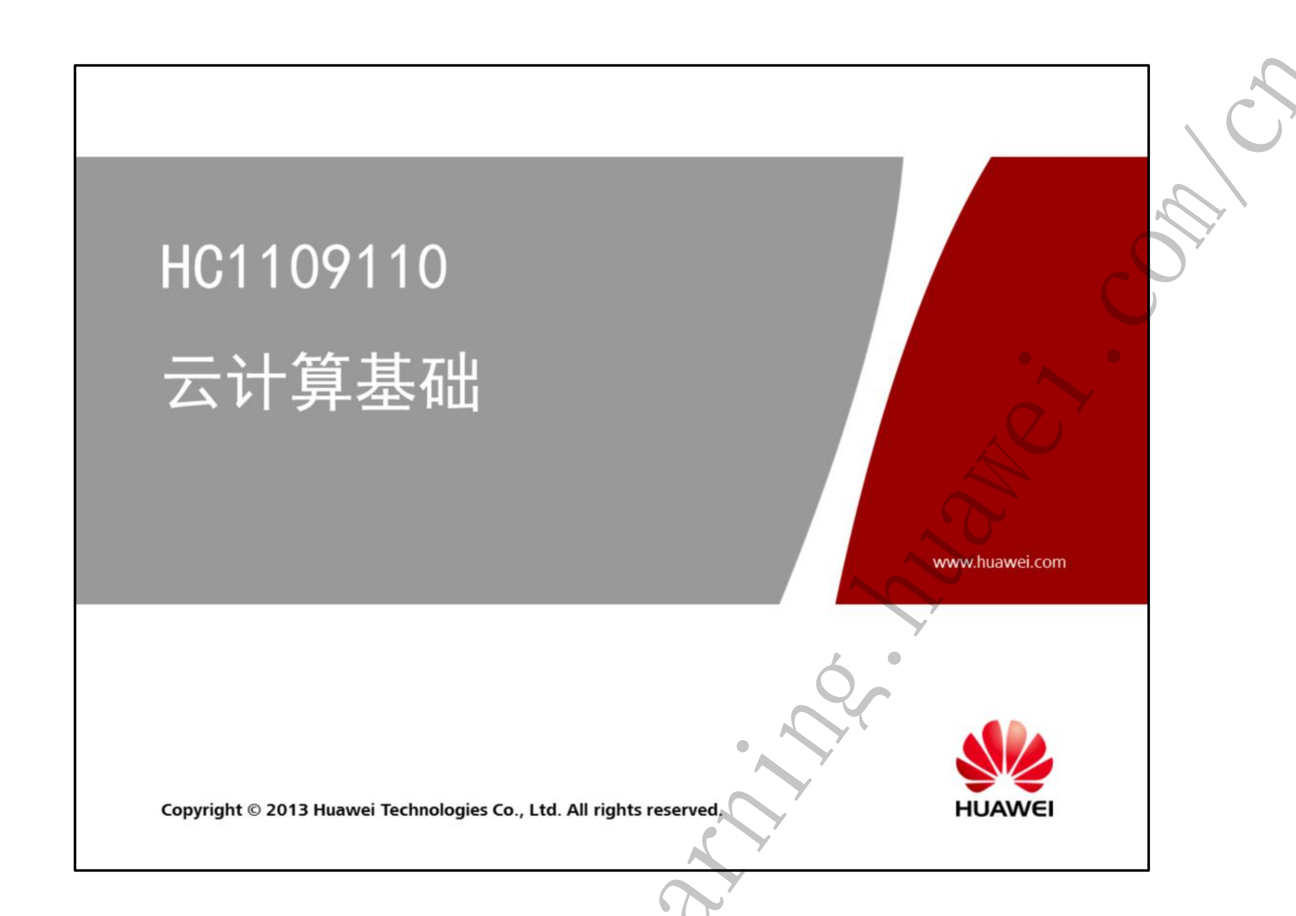
2、A B C

Thank you

www.huawei.com

HC1109110 云计算基础




HC1109110
云计算基础
www.huawei.com
Copyright © 2013 Huawei Technologies Co., Ltd. All rights reserved.
HUAWEI



目标

- 学完本课程后，您将能够:
  - 了解云计算的概念与背景
  - 掌握云计算的部署与商业模式
  - 了解云计算的核心技术与价值
  - 掌握华为云计算解决方案





目录



云计算的概念

- 云计算是一种通过Internet以服务的方式提供动态可伸缩的虚拟化资源的计算模式。(Cloud computing is a style of computing in which dynamically scalable and often virtualized resources are provided as a service over the Internet.)

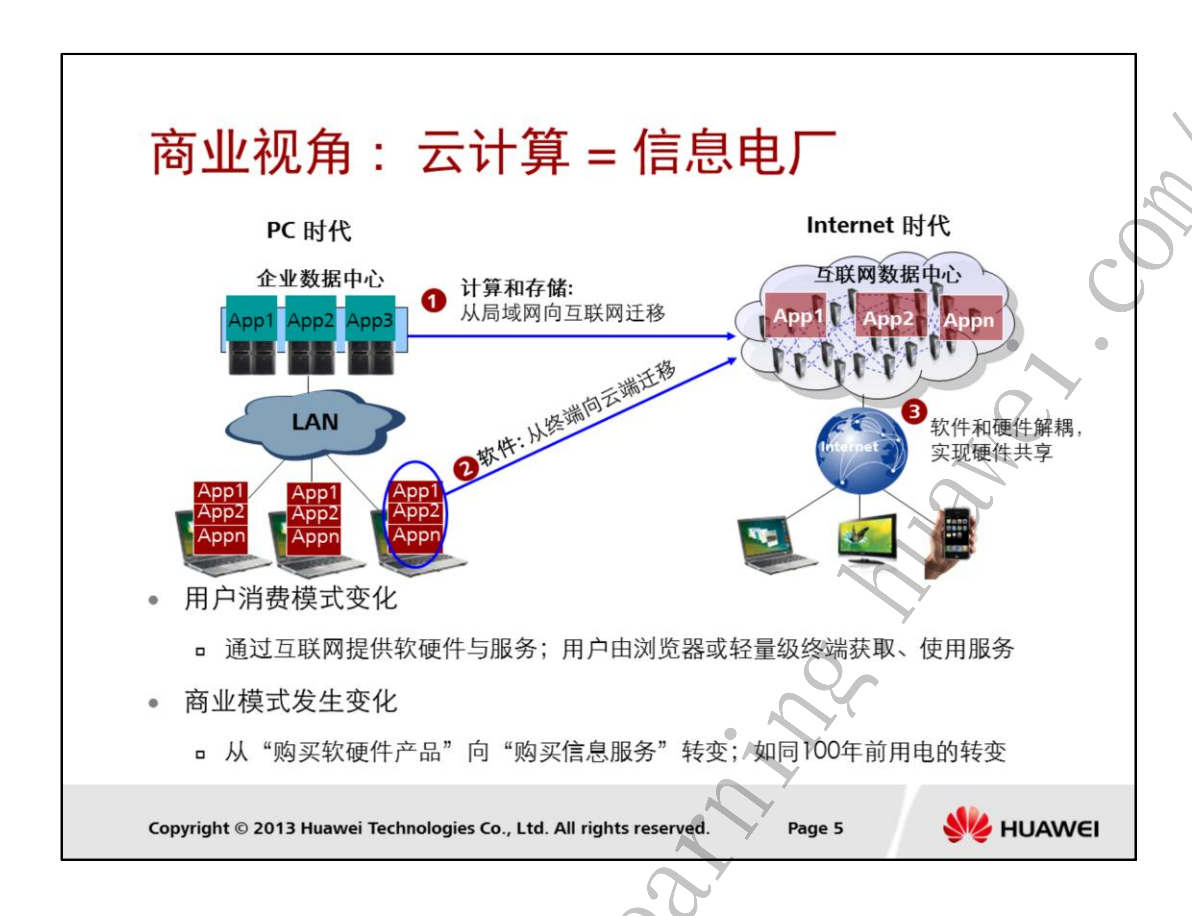
—— Wiki定义

 Page 4



云计算（cloud computing）是基于互联网的相关服务的增加、使用和交付模式，通常涉及通过互联网来提供动态易扩展且经常是虚拟化的资源。云是网络、互联网的一种比喻说法。过去在图中往往用云来表示电信网，后来也用来表示互联网和底层基础设施的抽象。狭义云计算指IT基础设施的交付和使用模式，指通过网络以按需、易扩展的方式获得所需资源；广义云计算指服务的交付和使用模式，指通过网络以按需、易扩展的方式获得所需服务。这种服务可以是IT和软件、互联网相关，也可是其他服务。它意味着计算能力也可作为一种商品通过互联网进行流通。

云计算的资源是动态易扩展而且虚拟化的，通过互联网提供。终端用户不需要了解“云”中基础设施的细节，不必具有相应的专业知识，也无需直接进行控制，只关注自己真正需要什么样的资源以及如何通过网络来得到相应的服务。

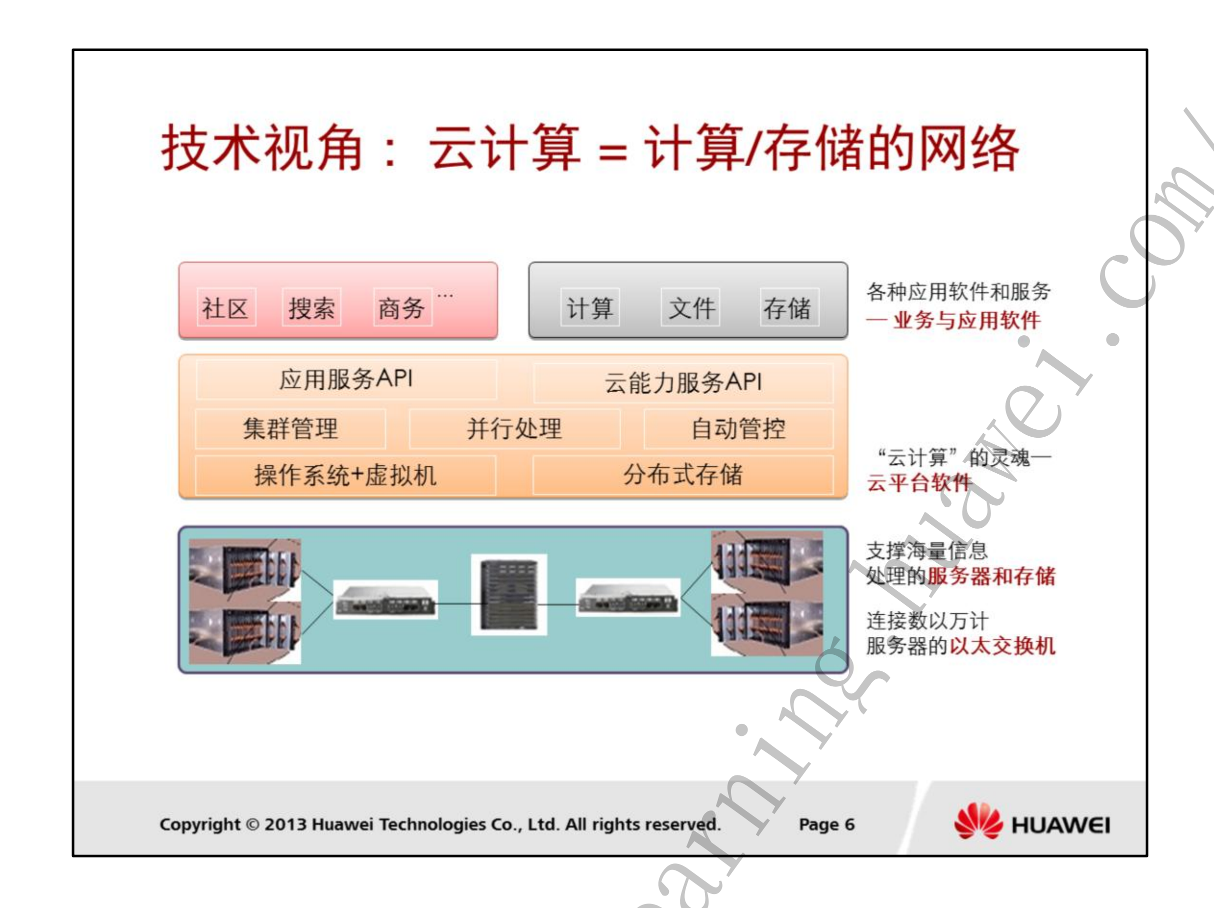


IT即服务，云计算就是建设信息电厂提供IT服务。

云计算是通过互联网提供软件、硬件与服务，并由网络浏览器或轻量级终端软件来获取和使用服务；即服务从局域网向Internet迁移，终端计算和存储向云端迁移；

从商业的视角看，云计算是以Google为代表的“Internet时代”和以微软/Intel为代表的“PC时代”之间两个时代的竞争；这个过程将是逐步的、长期的但呈加快的趋势；

云计算好比是从古老的单台发电机模式转向了电厂集中供电的模式。它意味着计算能力也可以作为一种商品进行流通，就像电一样，取用方便，费用低廉。



云计算与传统的单机和网络应用模式相比，具有如下特点：

虚拟化：这是云计算最核心的特点，包括资源虚拟化和应用虚拟化。每一个应用部署的环境和物理平台是没有关系的。通过虚拟平台进行管理达到对应用进行扩展、迁移、备份。

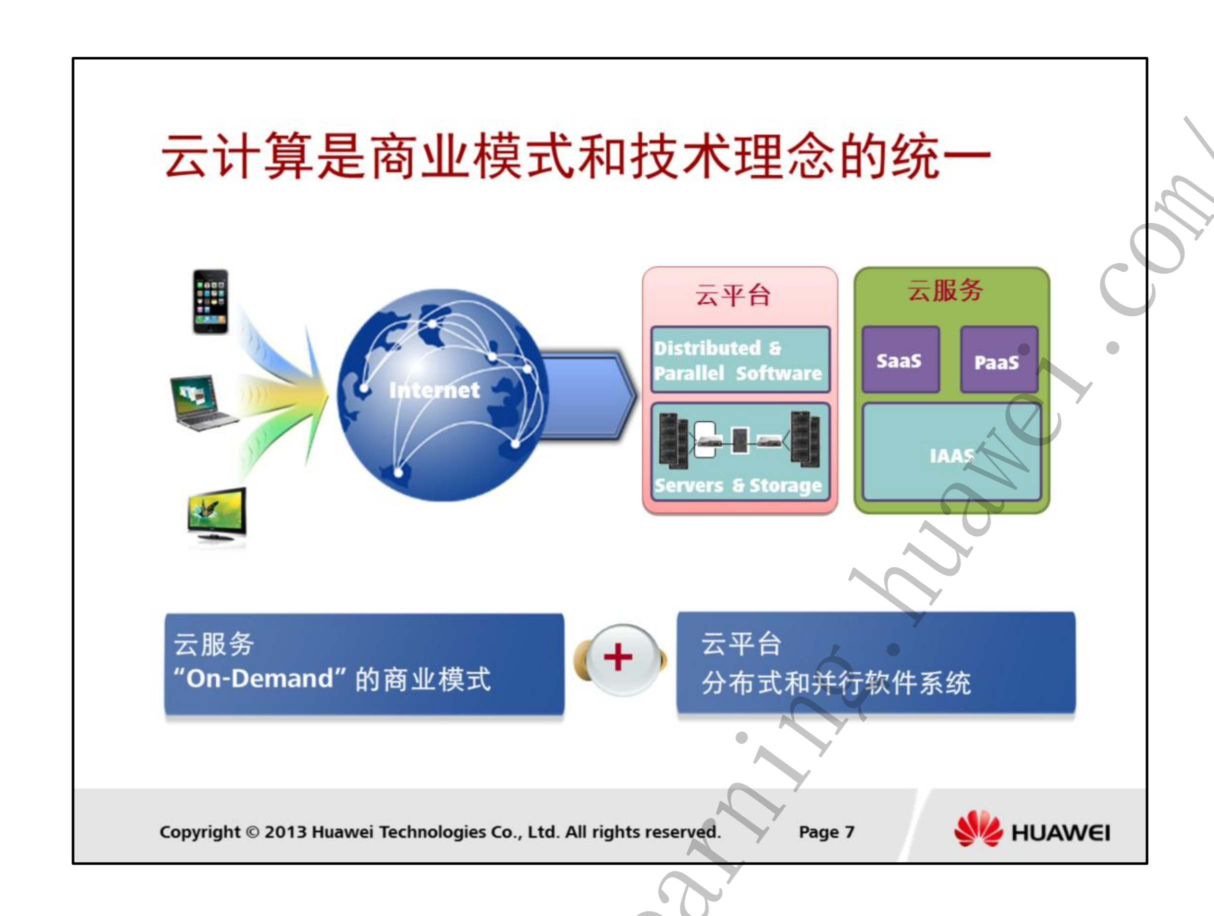
动态可扩展：通过动态扩展虚拟化的层次达到对应用进行扩展的目的。可以实时将服务器加入到现有的服务器集群中，增加计算能力。

按需部署：用户运行不同的应用需要不同的资源和计算能力。云计算平台可以按照用户的需求部署资源和计算能力。

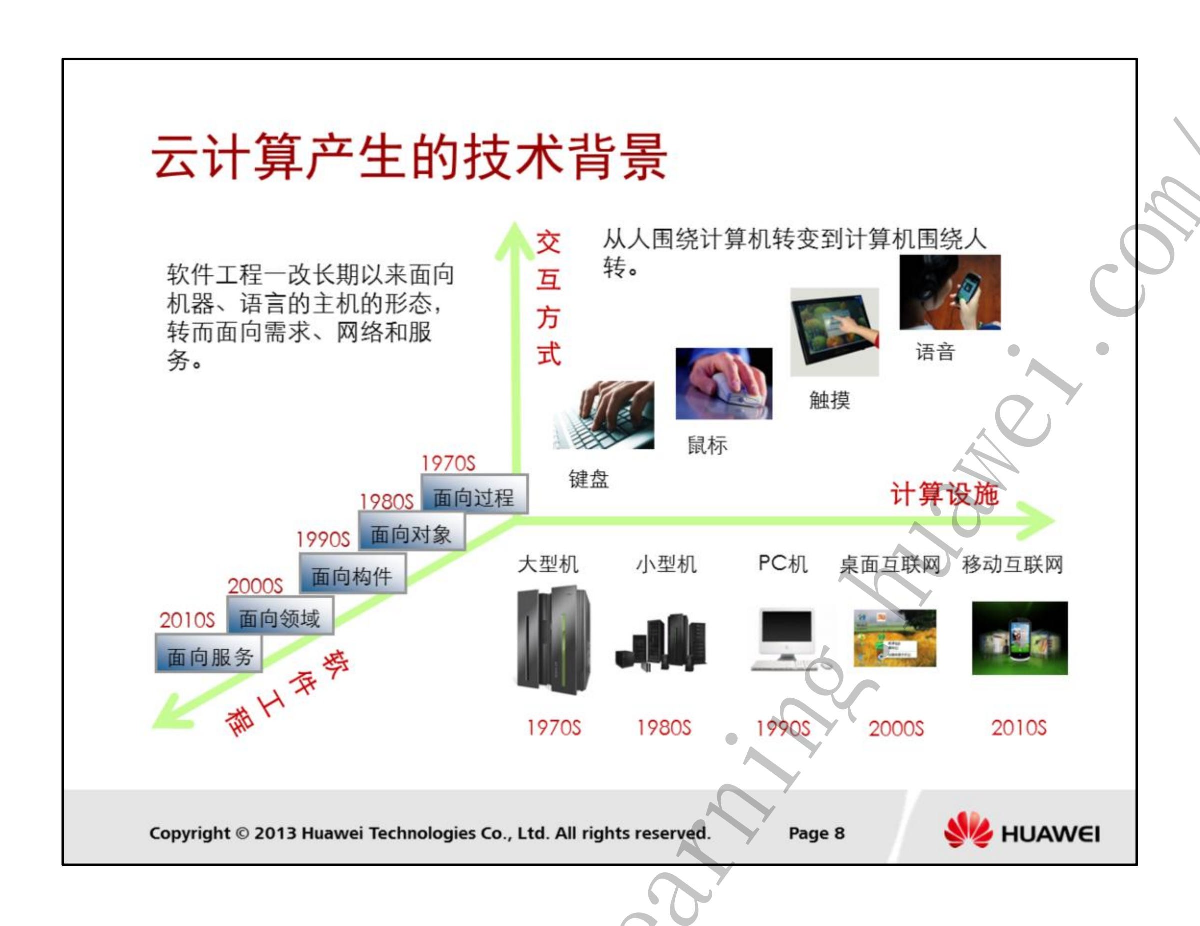
高灵活性：现在大部分的软件和硬件都对虚拟化有一定支持，各种IT资源（例如软件、硬件、操作系统、存储网络等）通过虚拟化，放在云计算虚拟资源池中进行统一管理。同时，能够兼容不同硬件厂商的产品，兼容低配置机器和外设而获得高性能计算。

高可靠性：虚拟化技术使得用户的应用和计算分布在不同的物理服务器上面，即使单点服务器崩溃，仍然可以通过动态扩展功能部署新的服务器，保证应用和计算的正常运转。

高性价比：云计算采用虚拟资源池的方法管理所有资源，对物理资源的要求较低。可以使用廉价的PC组成云，而计算性能却可超过大型主机。



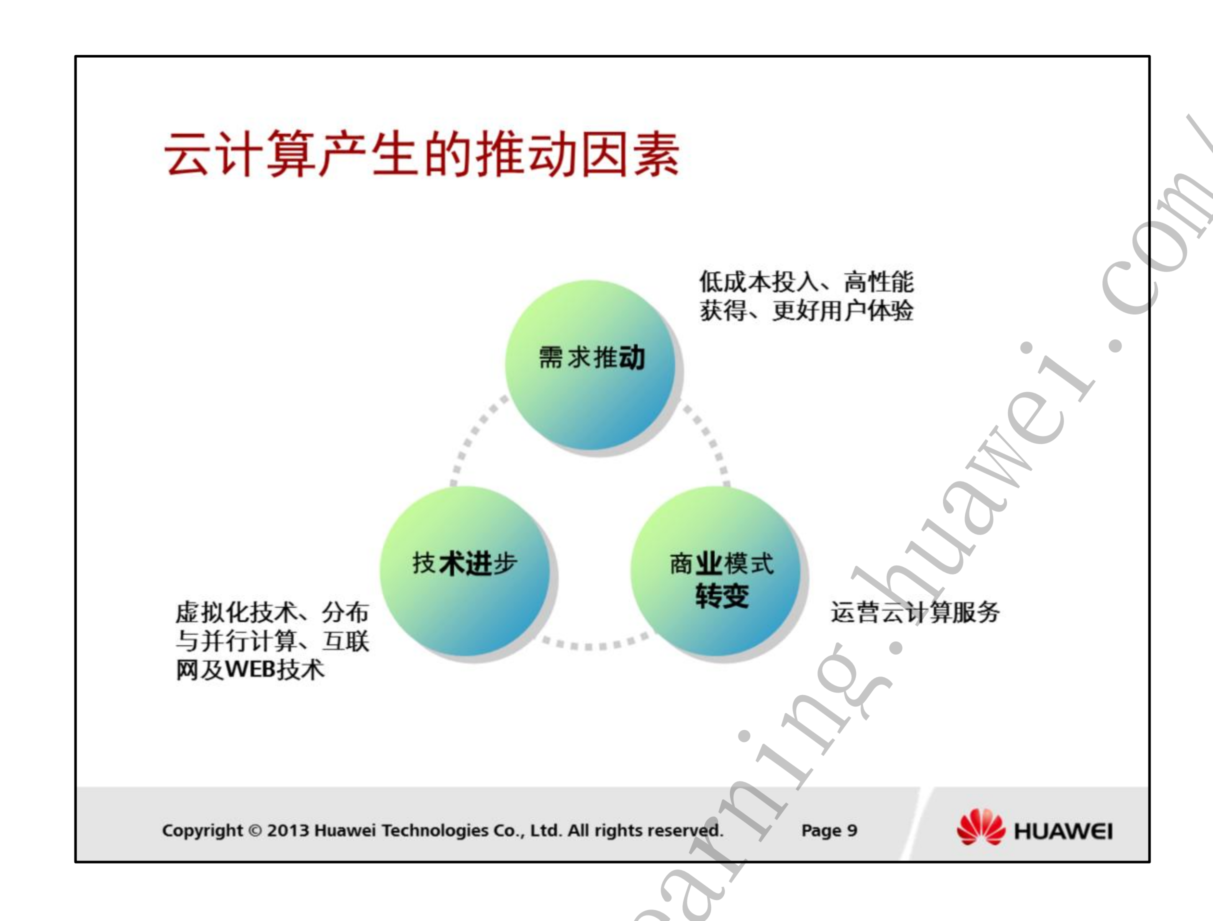
“On-Demand”商业模式，即按需服务模式。用户所需要的应用软件和数据都位于云中，用户使用客户端按照自己的需求访问所要的应用。云服务提供商同样按照客户的需求提供相应的服务，和收到相应费用。



云计算的产生，首先源于信息革命50多年来计算设施的变迁：从60年代的大型机，到70年代的小型机，到80年代的个人电脑和局域网，再到90年代对人类生产和生活产生了深刻影响的桌面互联网，以及目前大家所高度关注的移动互联网、无处不在的网络，计算设施不断地由单机向网络变迁。通信和网络的发展速度比以摩尔速度为代表的计算速度发展更快，这种变化是云计算能够生根发芽的土壤。

其次，软件工程40年来也发生了很大变化：70年代，人们把程序设计中的流程图看得很重要，80年代开始面向对象，90年代面向构件，现在我们面向领域、面向服务。软件工程一改长期以来面向机器、面向语言和面向中间件等面向主机的形态，转为面向需求和服务等面向网络的形态，真正实现了软件即服务，这是软件工程的大转身，服务将会成为云计算下软件开发的基本样式。

第三，半个世纪以来，人机交互方式也在逐渐发生改变，从主要以键盘的字符界面交互为主，发展到后来鼠标的图形界面，再到后来触摸、语音、手势等，各种各样便捷的交互方式使人围绕计算机转的时代已经过去，现在，计算机需要围着用户转，用户越来越只需要关注于核心业务，并不需要成为计算机或IT的“业余工程师”，这也正是云计算给用户在利用IT设施的方式上所带来的重要改变。

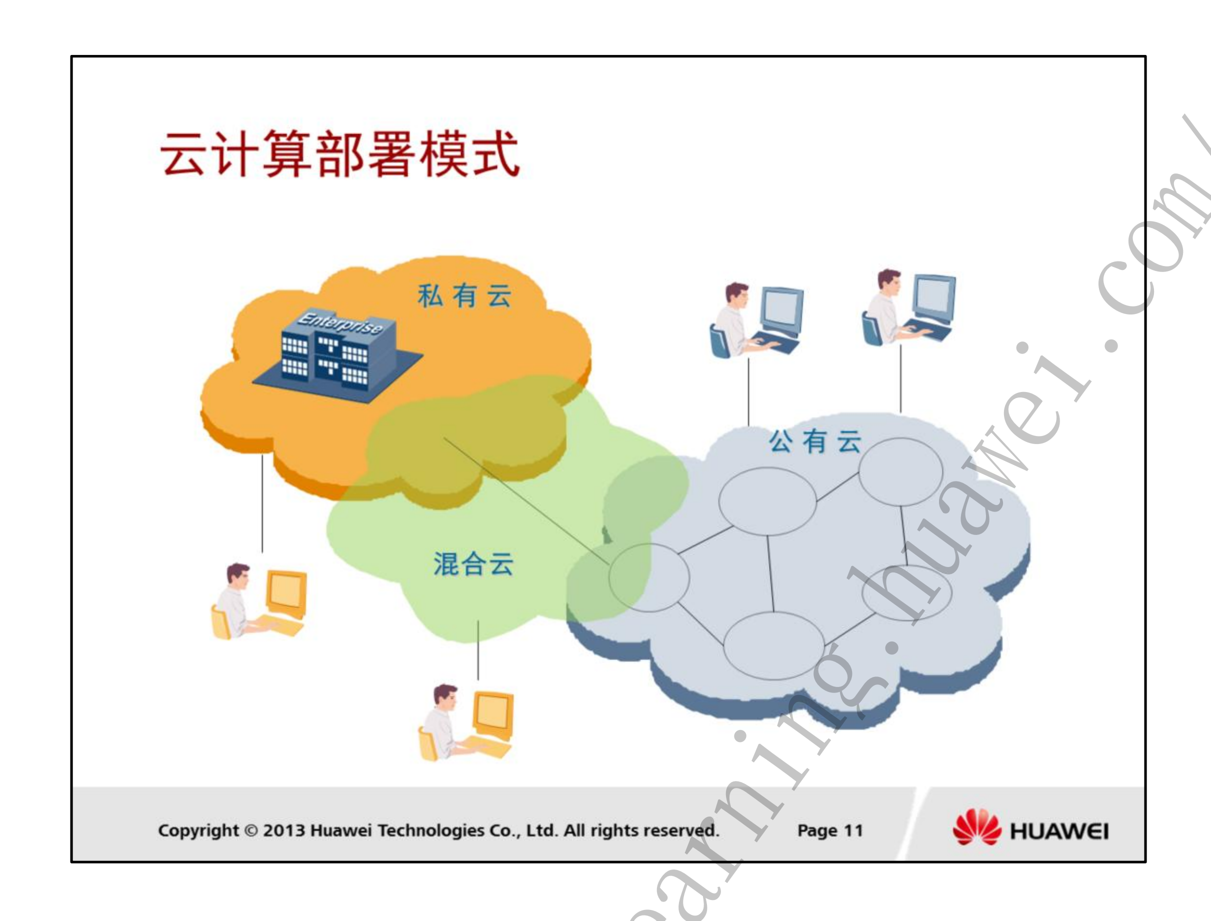


- 需求推动
  - 政企客户的信息化需求强烈，但同时要求低成本投入、高性能获得；
  - 个人用户的互联网、移动互联网应用需求强烈，追求更好用户体验。
- 技术进步
  - 虚拟化技术、分布与并行计算、互联网及WEB技术的发展与成熟，使得基于互联网提供包括IT基础设施、开发平台、软件应用成为可能；
  - 宽带技术及用户发展，使得基于互联网的服务使用模式逐渐成为主流。
- 商业模式转变
  - 少数云计算的先行者（例如Amazon的IaaS)的云计算服务已开始运营
  - 市场对云计算商业模式已有一定认可，越来越多的用户接受并使用云计算服务
  - 几年之内,云计算已从新兴技术发展成为当今的热点技术。从Google公开发布的核心文件到Amazon EC2（亚马逊弹性计算云）的商业化应用，再到美国电信巨头AT&T（美国电话电报公司）推出的Synaptic Hosting（动态托管）服务，云计算从节约成本的工具到盈利的推动器，从ISP（网络服务提供商）到电信企业，已然成功地从内置的IT系统演变成公共的服务。



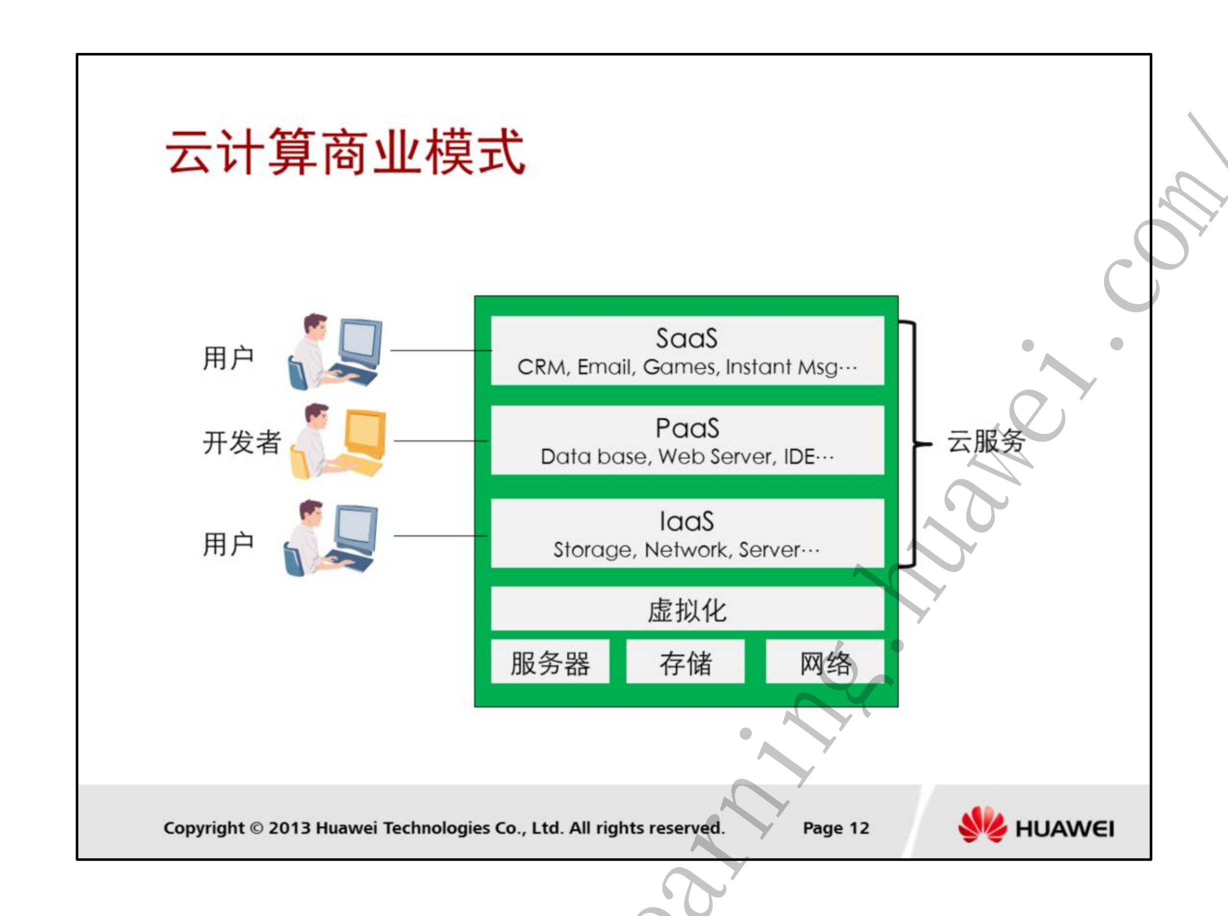
目录





云计算有三种部署模式：私有云计算、公有云计算、混合云计算。

- 私有云计算：一般由一个组织来使用，同时由这个组织来运营。华为数据中心属于这种模式，华为自己是运营者，也是它的使用者，也就是说使用者和运营者是一体，这就是私有云。
- 公有云计算：就如共用的交换机一样，电信运营商去运营这个交换机，但是它的用户可能是普通的大众，这就是公共云。
- 混合云计算：它强调基础设施是由二种或更多的云来组成的，但对外呈现的是一个完整的实体。企业正常运营时，把重要数据保存在自己的私有云里（比如：财务数据），把不重要的信息放到公有云里，两种云组合形成一个整体，就是混合云。比如说电子商务网站，平时业务量比较稳定，自己购买服务器搭建私有云运营，但到了圣诞节促销的时候，业务量非常大，就从运营商的公有云租用服务器，来分担节日的高负荷；但是可以统一的调度这些资源，这样就构成了一个混合云。



- IaaS（Infrastructure as a service）

基础设施即服务，提供给消费者的服务是对所有基础设施的使用，包括计算、存储、网络和其它的计算资源，用户能够部署和运行任意软件，包括操作系统和应用程序。消费者不需要管理或控制任何云计算基础设施，但能控制操作系统的选择、存储空间、部署的应用，也有可能获得网络组件（例如，防火墙，负载均衡器等）的控制。Amazon EC2是IaaS的典型代表。

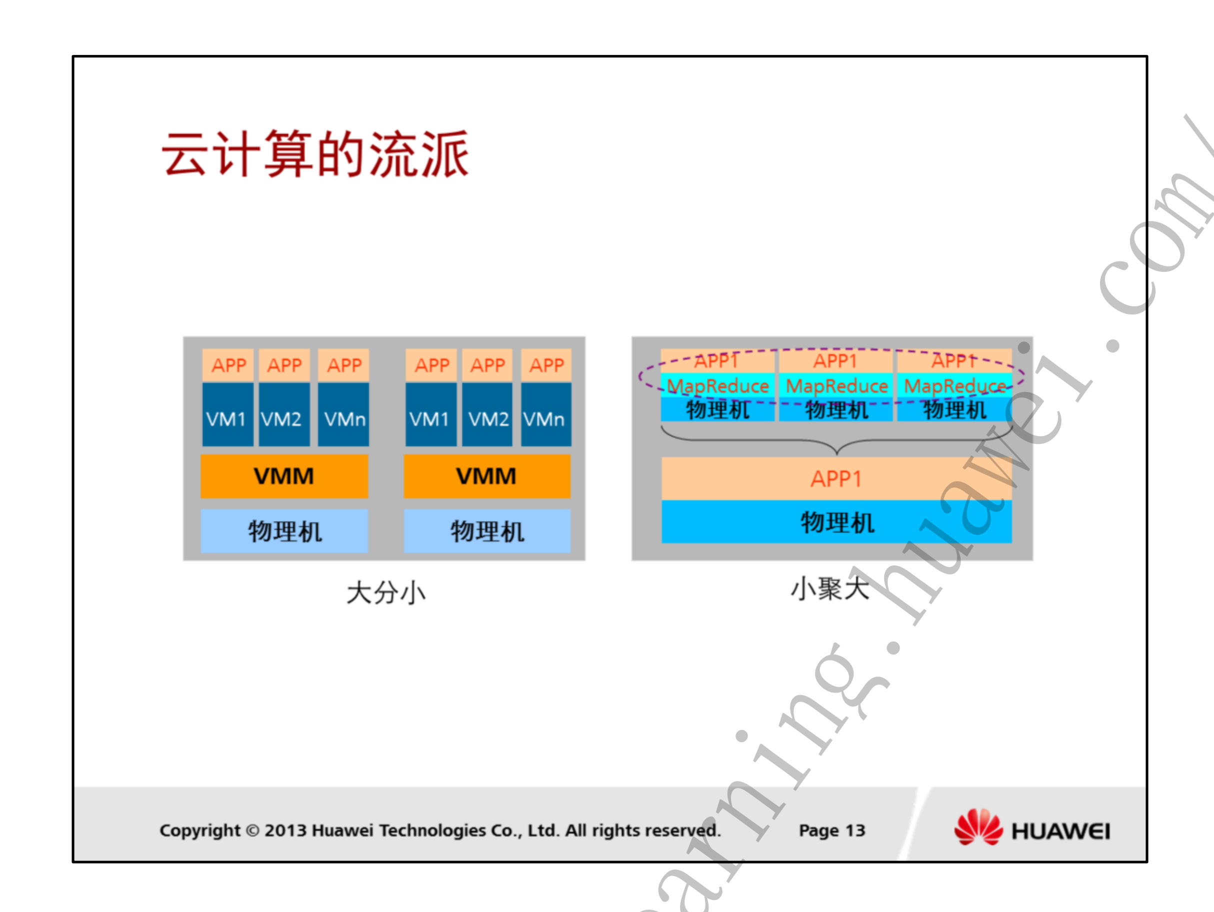
- PaaS（Platform as a service）

平台即服务，提供给消费者的服务是部署在云计算基础设施上的应用程序开发平台（例如Java，.Net等）。客户不需要管理或控制底层的云计算基础设施，但能控制部署的应用程序开发平台。Microsoft Azure是PaaS的典型代表。

- SaaS（Software as a service）

软件即服务，提供给消费者的服务是运行在云计算基础设施上的应用程序，如CRM、ERP、OA等。Salesforce online CRM就属于SaaS。

除此之外，还有另外一种模式：DaaS，即Data as a service，是一种以数据存储为服务内容的服务模式。



在业界，云计算的部署形态主要有两种，一种是大分小，另一种是小聚大。

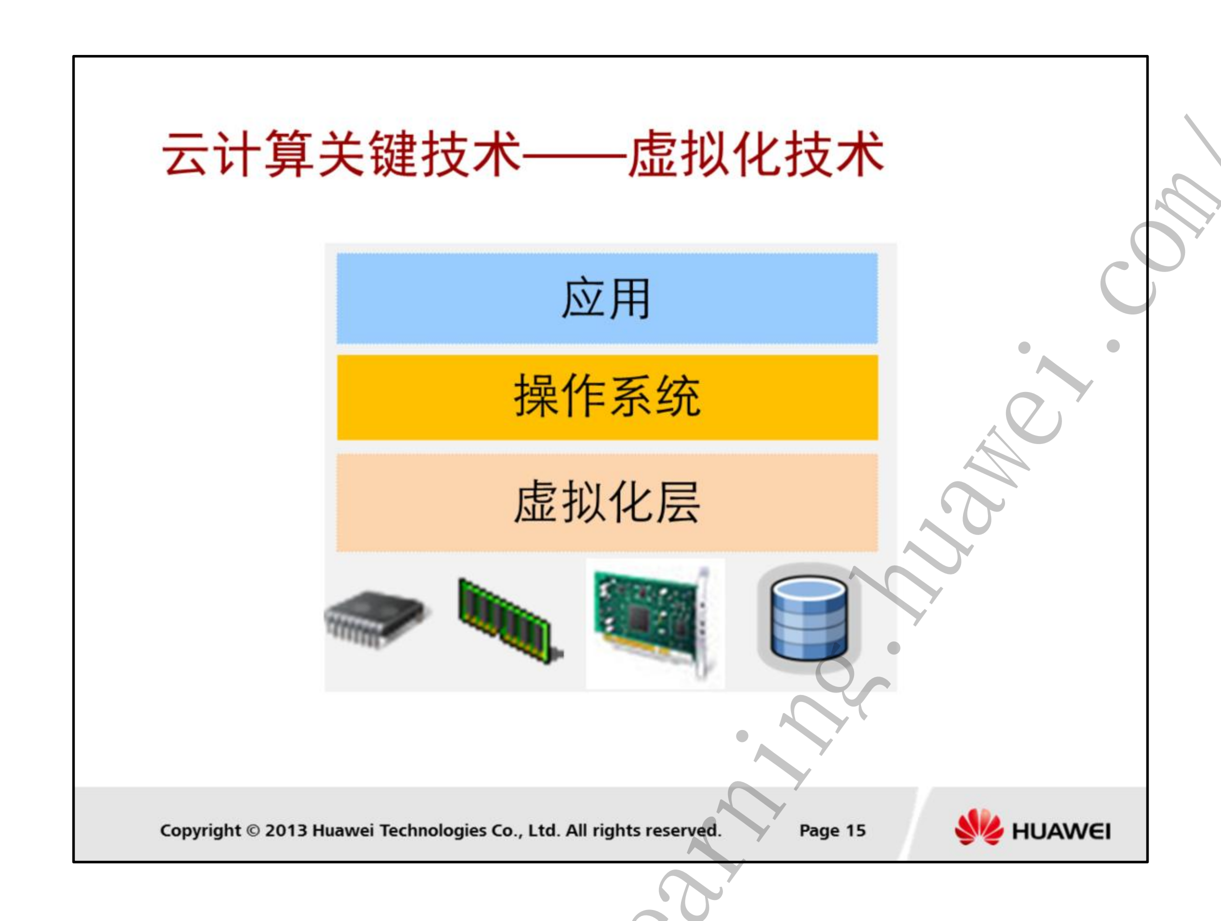
大分小是指通过虚拟化技术，对性能强大的物理机资源进行虚拟化，形成可动态调整分配的计算、存储、网络资源池。这种方式的关键技术主要包括虚拟化、虚拟机监控、虚拟机调度与迁移，主要适用于资源可时分复用的场景。Amazon EC2业务是大分小形式的典型代表。

小聚大指的是，把多个性能较低的物理资源，通过一定的软硬件方式，形成逻辑上性能强大的物理资源。一个对资源需要较大的任务，可以分配到各个小的物理机进行处理。这种方式主要涉及任务分解与调度、分布式通信总线、全局一致性等技术。Google是小聚大的典型厂家。

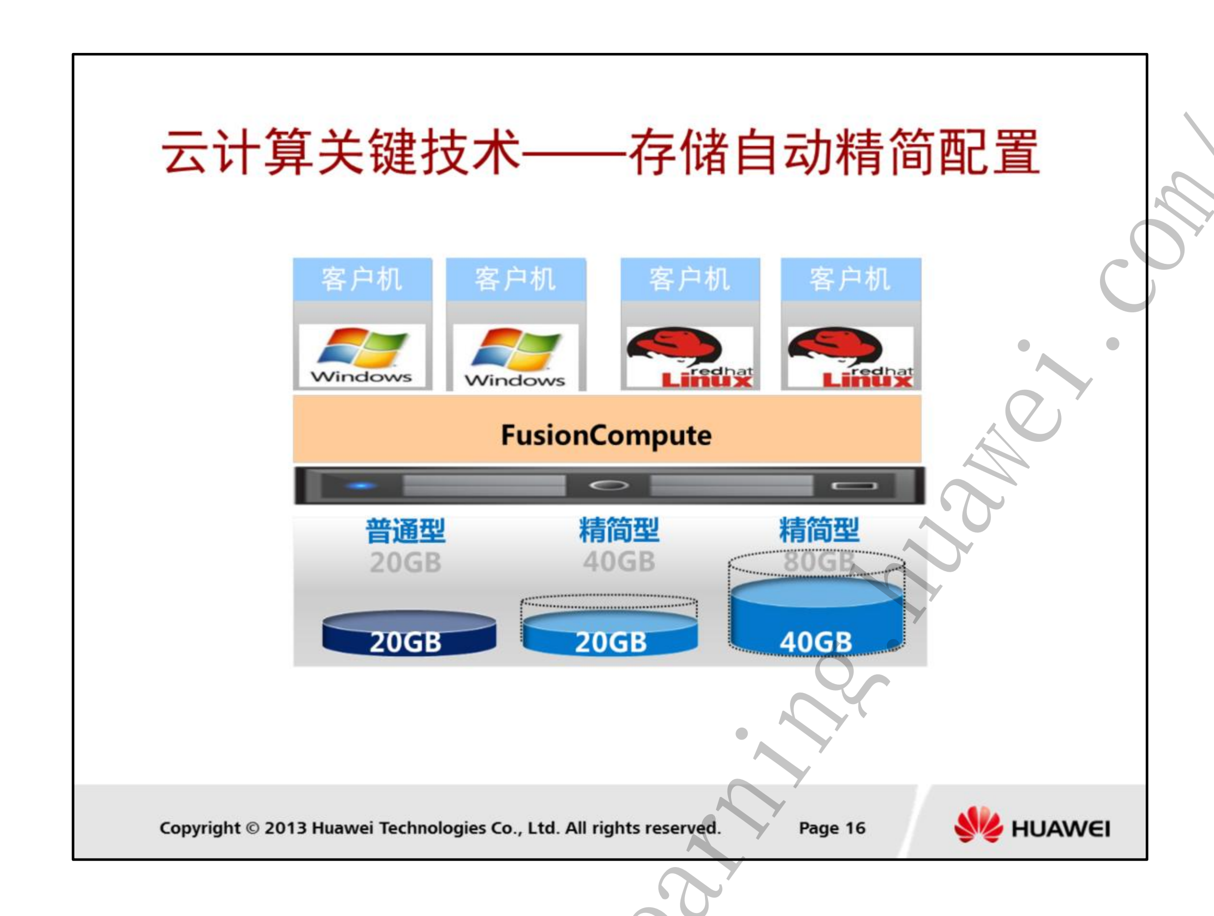


目录





虚拟化实现了IT资源的逻辑抽象和统一表示，在大规模数据中心管理和解决方案交付方面发挥着巨大的作用，是支撑云计算伟大构想的最重要的技术基石。采用虚拟化技术对计算、存储、网络等物理资源进行虚拟化，形成资源池对用户提供按需服务。

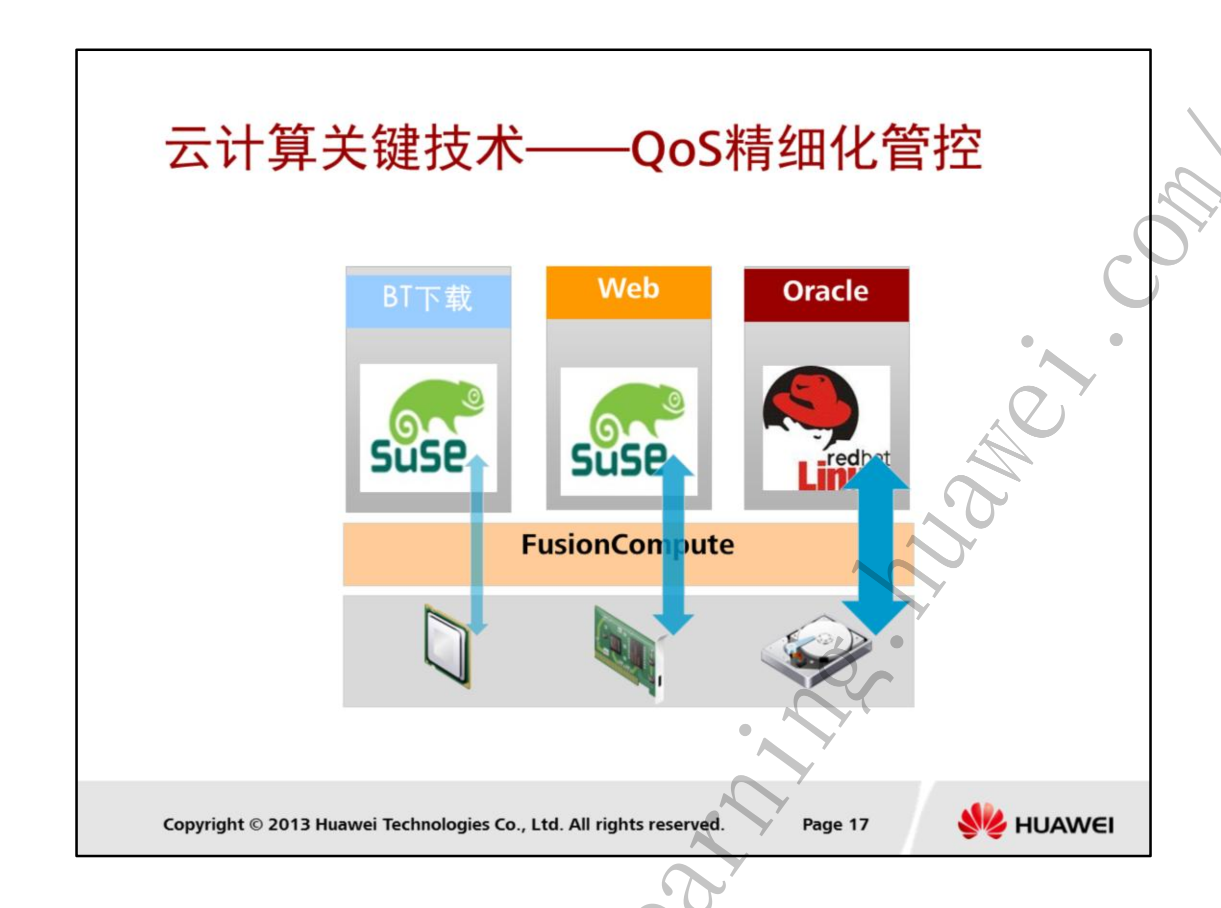


虚拟存储精简置备是一种通过以灵活的按需方式分配存储空间来优化存储利用率的方法。精简置备与传统模式（称为厚置备）截然不同：厚置备预先提供大量存储空间以满足未来的存储需要，但是空间可能一直未被使用，这样会导致无法充分利用存储容量；精简置备可以为客户虚拟出比实际物理存储更大的存储空间，只有写入数据的存储空间才会为之真正分配物理存储，未写入的虚拟存储空间不占用物理存储资源，从而提高存储利用率。虚拟存储精简配置基于虚拟磁盘级别提供，管理员可以按“普通”格式或“精简”格式分配虚拟磁盘文件。通过将虚拟磁盘设为“精简”配置，虚拟存储管理为虚拟磁盘在数据存储上分配存储空间，在开始时仅交付存储数据所需的存储空间，并不交付已经分配但并未使用的存储空间，并且随着虚拟磁盘上数据量的增加而增加空间供应量。当所有分配空间都已经交付时，“精简”磁盘与“普通”磁盘没有差异。

存储无关:虚拟存储精简配置与操作系统和硬件完全无关，因此只要使用虚拟镜像管理系统，就能提供虚拟存储精简配置功能。

容量监控:提供数据存储容量预警，可以设置阈值，当存储容量超过阈值时发出告警。

空间回收：提供虚拟磁盘空间监控和回收功能，当虚拟磁盘使用了一段时间，分配空间较大而实际使用空间较小时，可以通过磁盘空间回收功能回收已经分配交付但不在使用的空间。当前支持NTFS格式的虚拟机磁盘回收。

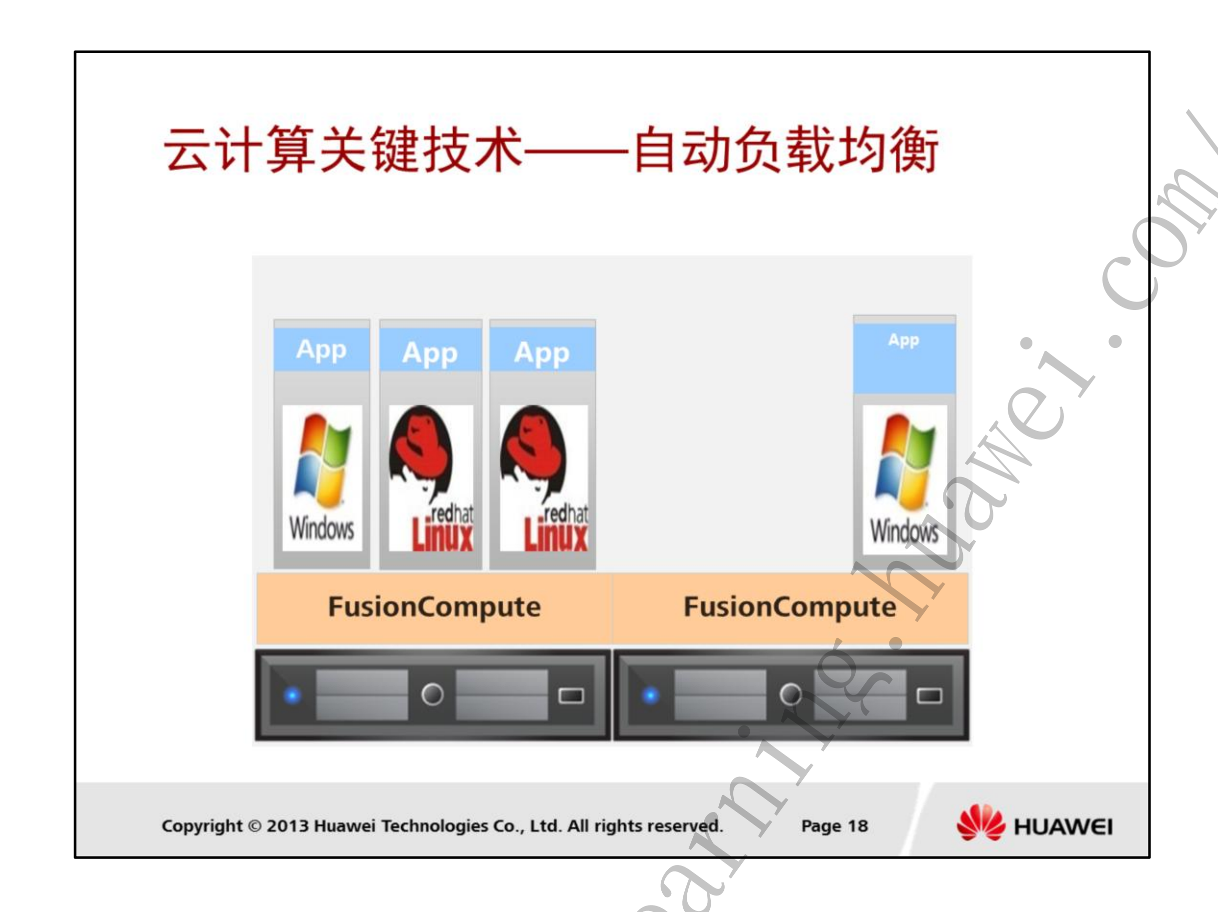


- CPU QoS
  - 预留百分比：控制虚拟机获得的最低计算能力，其值为虚拟机CPU配置值与预留量的乘积。
  - 限制百分比：控制虚拟机获得的最大计算能力，其值为虚拟机CPU配置值与限制量的乘积。
- 内存QoS

支持设置预留百分比：控制虚拟机最低可获取的物理内存，其值为虚拟机内存配置值与预留量的乘积。

- 网络QoS

采用虚拟交换机支持网络QoS（带宽限速及优先级），采用智能网卡实现虚拟接口最大带宽的速率限制。



动态资源调度(DRS)：动态分配和平衡资源，采用智能调度算法，根据系统的负载情况，对资源进行智能调度，达到系统的负载均衡，保证系统良好的用户体验。

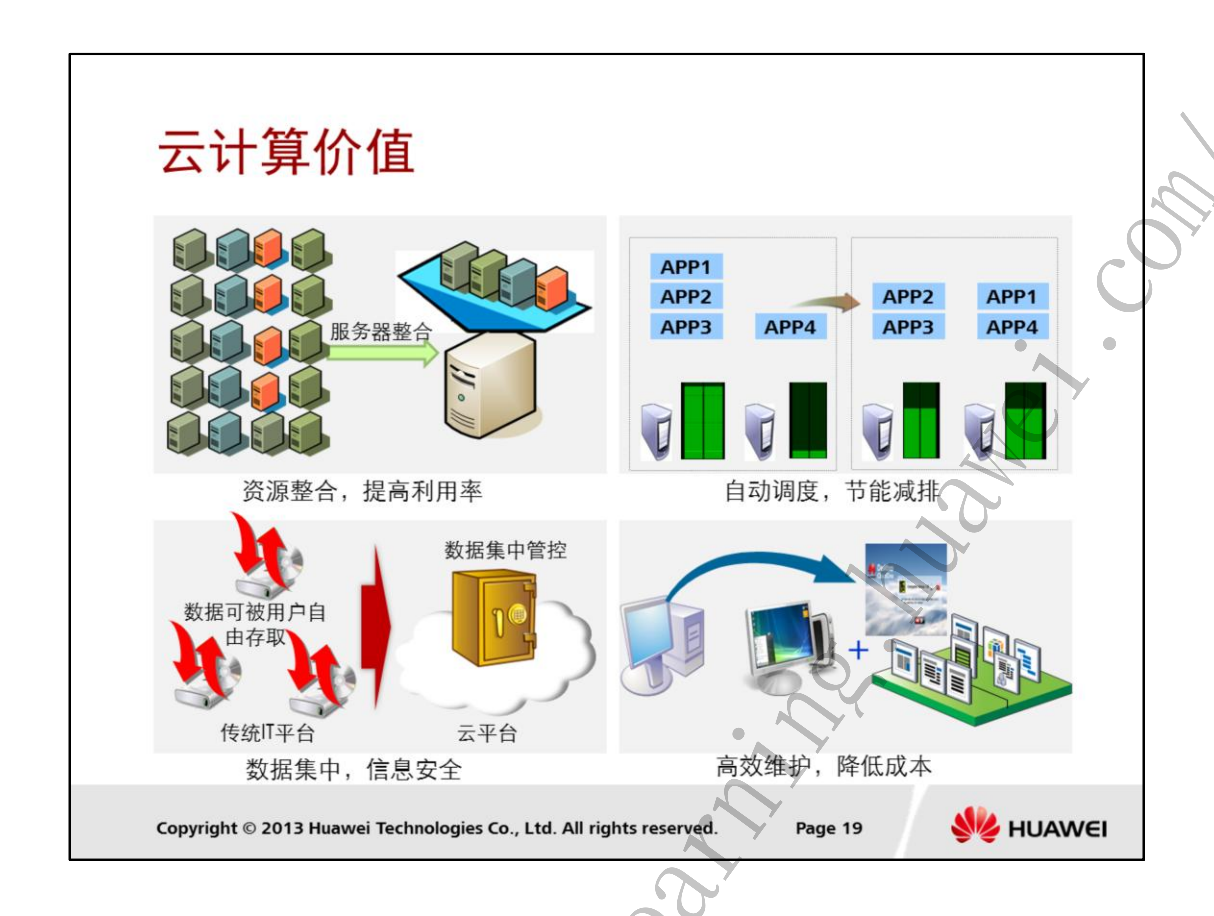
- 功能描述

动态资源调度策略针对集群 (Cluster) 设置，支持负载均衡策略。

负载均衡调度策略中，可以设置调度阈值、定义策略生效的时间段。在策略生效的时间段内，如果某主机的CPU、内存负载阈值超过调度阈值，系统就会自动迁移一部分虚拟机到其它CPU、内存负载低的主机中，保证主机的CPU、内存负载处于均衡状态。

- 应用场景

本特性应用于计算资源的负载均衡和节能减排场景。



- 资源整合提高利用率

利用虚拟化技术，实现资源的弹性伸缩。每台服务器虚拟出多台虚拟机，避免原来的服务器只能给某个业务独占的问题；可通过灵活调整虚拟机的规格（CPU、内存等），增加虚拟机/减少虚拟机，快速满足业务对计算资源需求量的变化；利用虚拟化计算，将一定量的物理内存资源虚拟出更多的虚拟内存资源，可以创建更多的虚拟机。

- 自动调度，节能减排

基于策略的智能化、自动化资源调度，实现资源的按需取用和负载均衡，消峰填谷，达到节能减排的效果，白天基于负载策略进行资源监控，自动负载均衡，实现高效热管理；夜晚基于时间策略进行负载整合，将不需要的服务器关机，最大限度降低耗电量。

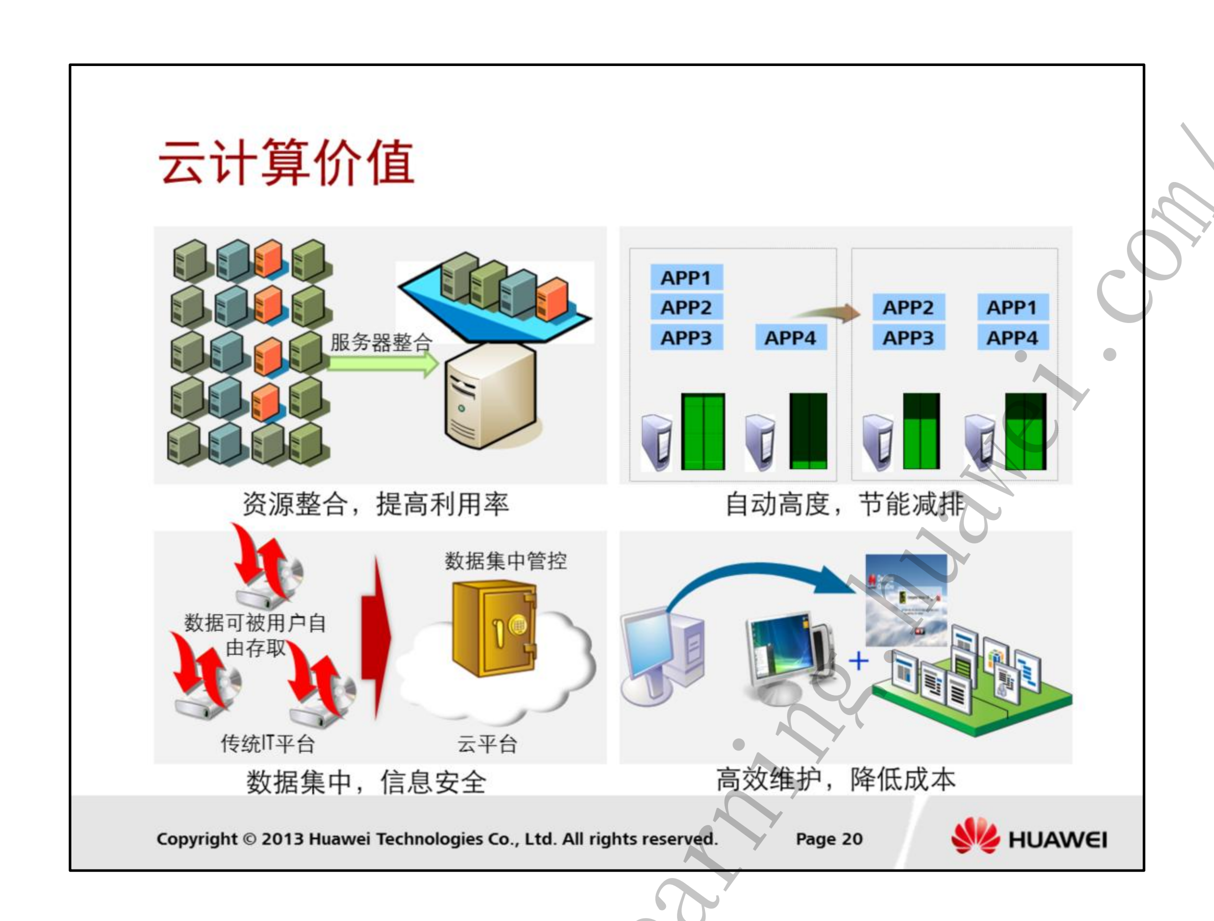
节能减排即动态电源管理(DPM)可以优化数据中心的能耗。开启DPM后，当集群中虚拟机使用资源比较低时，可以聚合虚拟机到少量主机，并关闭其它无虚拟机运行的主机，实现节能减排。当虚拟机所需资源增加时，DPM动态上电主机，确保提供足够资源。

- 数据集中，信息安全

传统IT平台，数据分散在各个业务服务器上，可能存在某单点有安全漏洞的情况。部署云系统后，所有数据集中在系统内存放和维护，并提供以下网络传输、系统接入、数据安全、趋势防病毒软件等安全保障。

- 高效维护，降低成本

对于一个大型企业，传统办公设备是人手一台PC机，如果PC发生故障，需要IT人员对发生故障的PC逐一进行维护。从故障报修，到IT故障排除将耗费大量人力和时间。而应用云

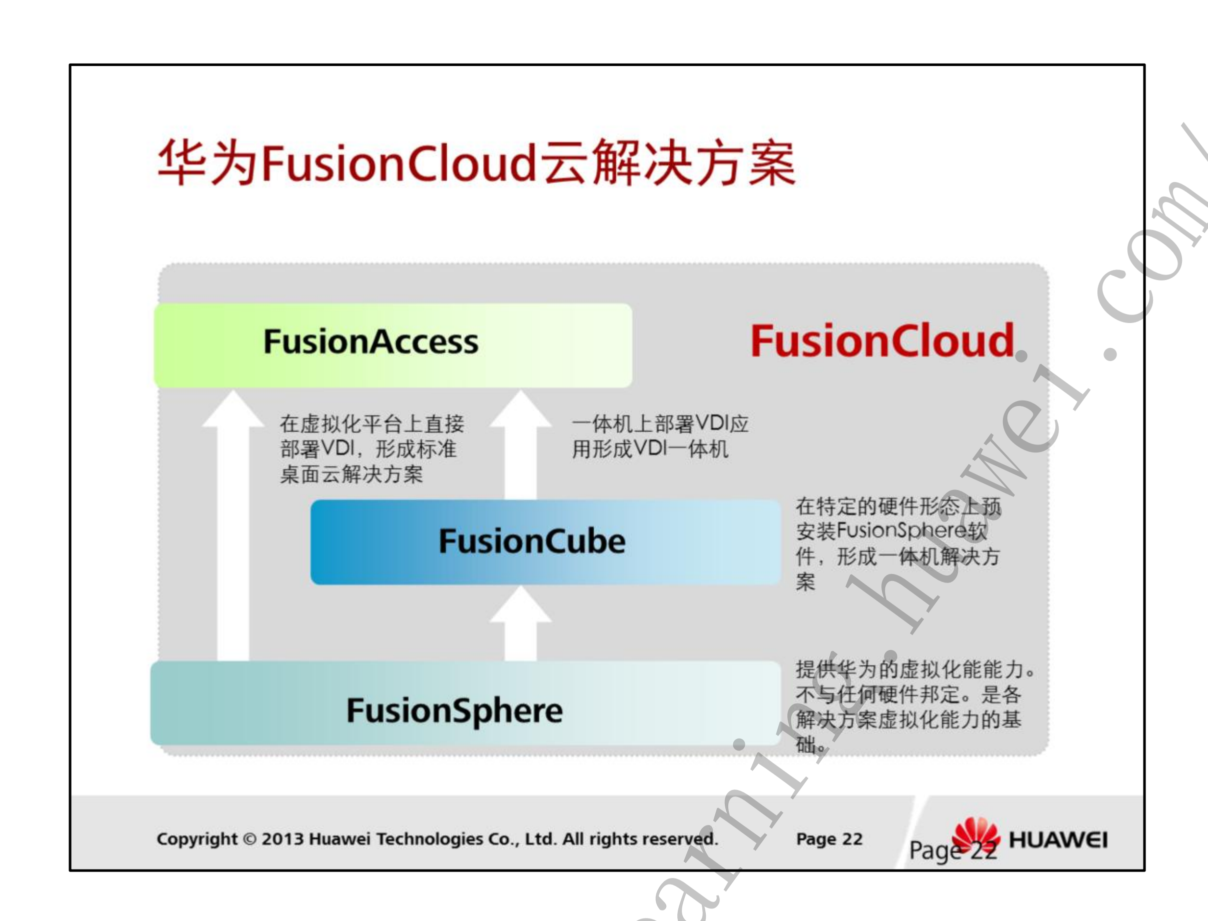


桌面办公方式，员工的办公系统和数据都在云端，前端免维护，相比传统维护方式将节省大量人力和时间。

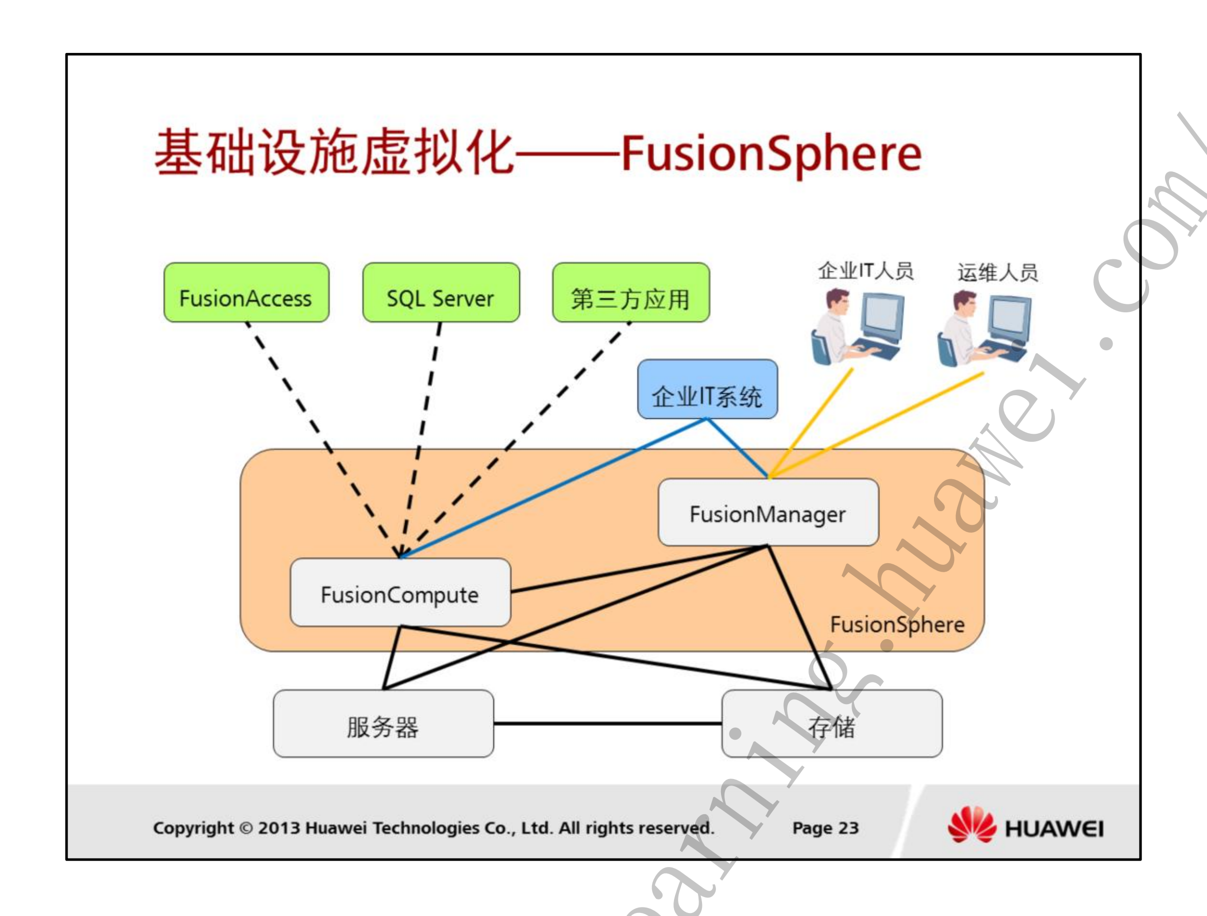


目录





华为目前可提供FusionSphere(基础设施虚拟化)、FusionCube(一体机)、FussionAccess(桌面云)三种云计算解决方案。FusionSphere基础设施虚拟解决方案是其它解决方案的基础，实现对物理基础设施的虚拟化。在特定硬件形态上预装FusionSphere软件可形成FusionCube一体机解决方案，实现业务快速部署。在FusionCube或FusionSphere的基础上部署VDI(Virtual Desktop Infrastructure)，可形成FusionAccess桌面云解决方案。



FusionSphere解决方案通过在服务器上部署虚拟化软件，将硬件资源虚拟化，从而使一台物理服务器可以承担多台服务器的工作。通过整合现有的工作负载并利用剩余的服务器以部署新的应用程序和解决方案，可以实现较高的整合率。

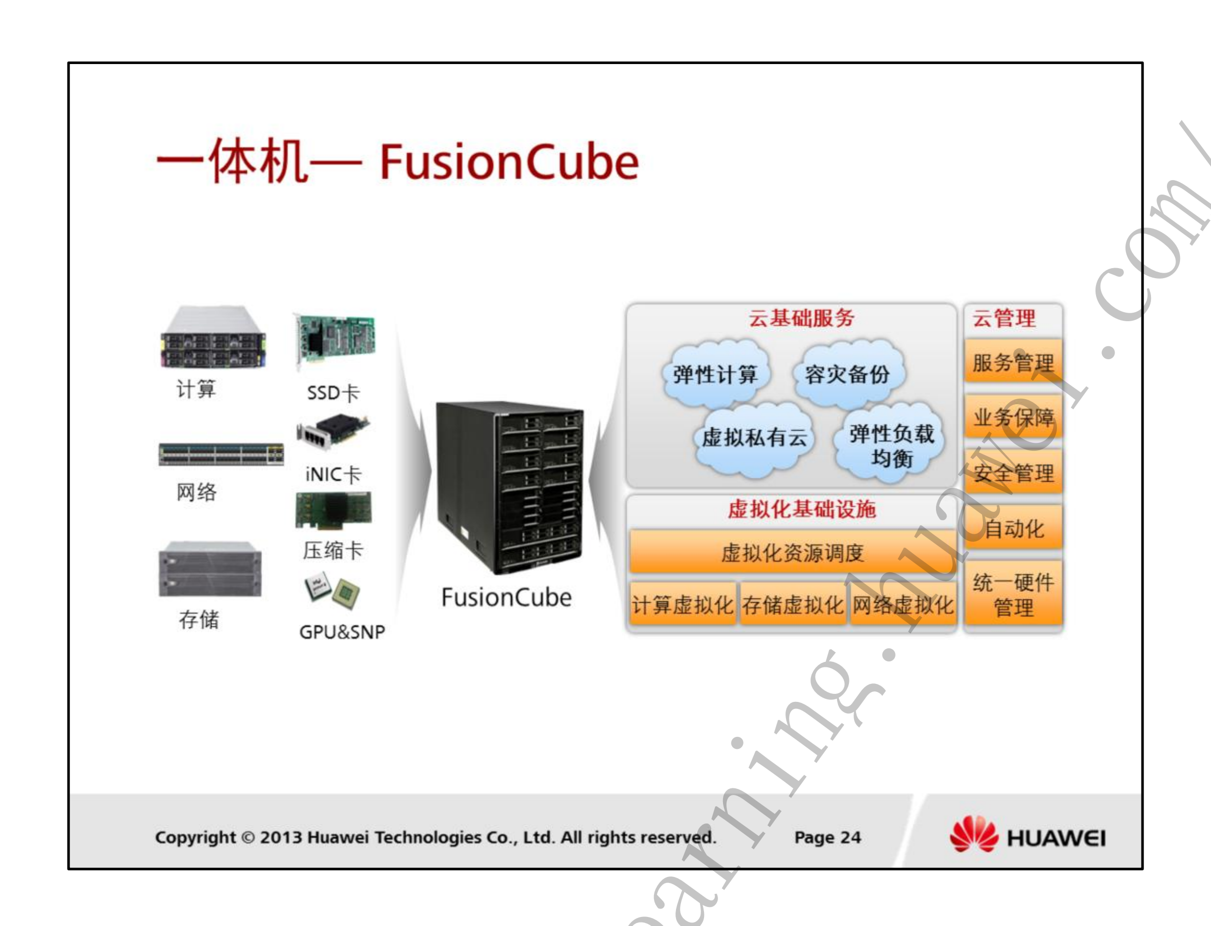
解决方案包括FusionCompute和FusionManager两软件组件。

FusionCompute软件组件主要包括VRM、主机组件，实现虚拟化物理资源，向数据中心虚拟化提供虚拟机服务。

FusionManager软件组件，主要包括IRM（Integrated Resource Management 集成资源管理）、SSP(Self-service Provisioning 自动发放)、AME(Automatic Management Engine 自动运维)、IAM(Identity and Access Management 身份和访问管理)、Uportal(Unified Portal 统一Portal)、IDB（intelligent data base）、CSB(Common Service and Bus 公共服务与总线)、UHM((unified hardware management统一硬件管理)等组件。作为数据中心虚拟化的管理软件，管理虚拟化资源、硬件资源，并提供服务管理等功能。

FusionManager通过SNMP接口向上级网管上报告警，计算、存储和网络设备可以通过SNMP、IPMI或SSH接入FusionManager。

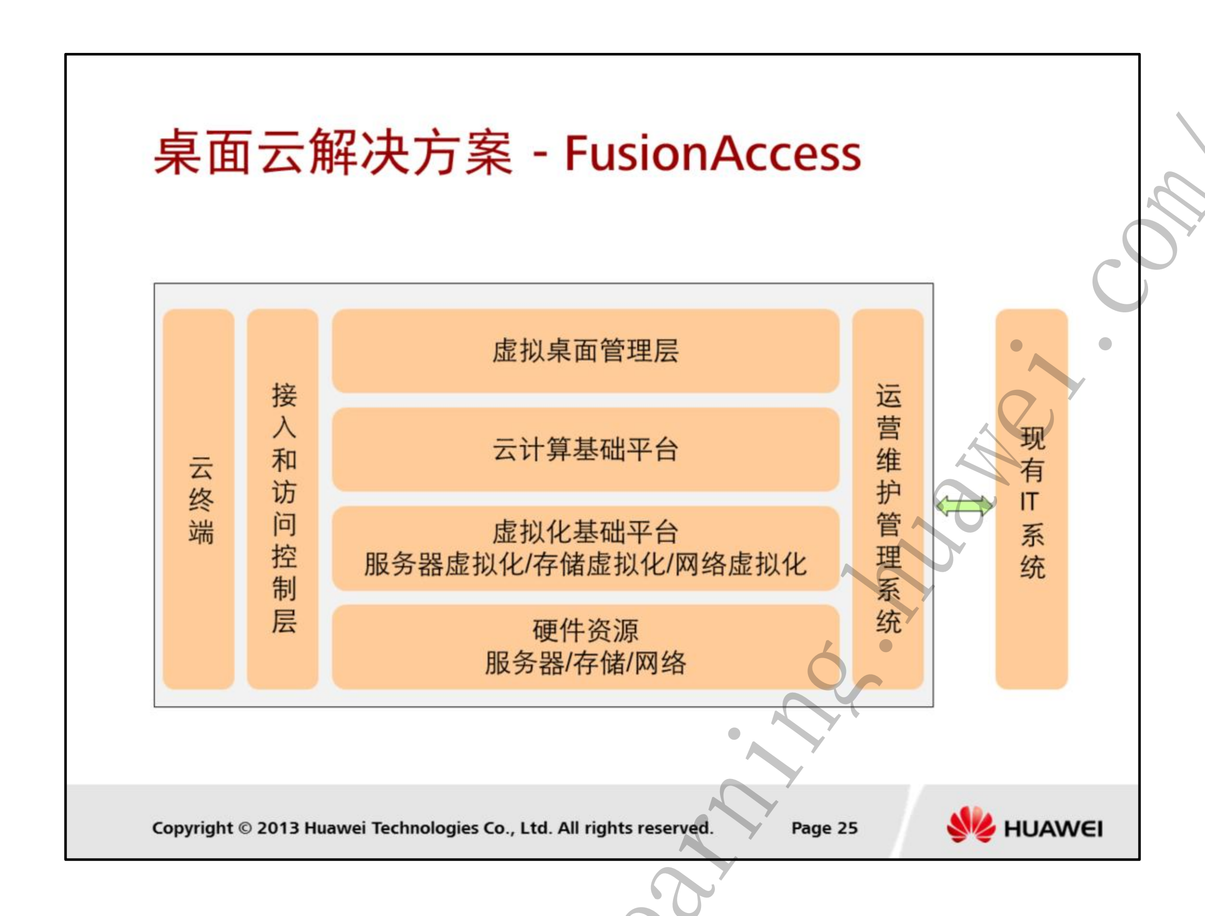
同时，FusionManager从虚拟化软件FusionCompute获取虚拟化资源配置、告警等信息；虚拟化软件根据FusionManager的指令，对虚拟机进行管理。



FusionCube一体机解决方案，硬件上集计算、存储、交换于一体，通过预装FusionCompute、FusionManage、FusionStorage软件，实现硬件资源的虚拟化和统一管理。

FusionCube作为虚拟化一体机，是一个开放的、可扩展的系统。通过FusionCube的统一资源管理、应用自动部署等特性，帮助用户简单、快速实现不同云应用的部署和维护管理。

用户可根据其业务需求，在FusionCube上定制开发、部署、更新或管理业务应用。这些应用包括简单的单机应用、复杂的集群应用等，如Exchange、SharePoint、ERP（Enterprise Resource Planning）/CRM（Customer Relationship Management）、VDI（Virtual Desktop Infrastructure）、SQL Server等。



华为桌面云解决方案是基于华为FusionCube或FusionSphere的一种虚拟桌面应用，通过在云平台上部署软、硬件，使终端用户通过瘦客户端（TC）或者其他任何与网络相连的设备来访问跨平台的应用程序，以及整个客户桌面。

华为桌面云解决方案重点解决传统PC办公模式给客户带来的如：安全、投资、办公效率等方面的诸多挑战，适合金融、大中型企事业单位、政府、呼叫中心、营业厅、医疗机构、军队或其他分散/户外/移动型办公单位。

华为桌面桌面云解决方案逻辑架构包括：

- 硬件资源

提供部署桌面云系统相关的硬件基础设施，包括服务器、存储设备、交换设备、机架、安全设备、防火墙、配电设备等。

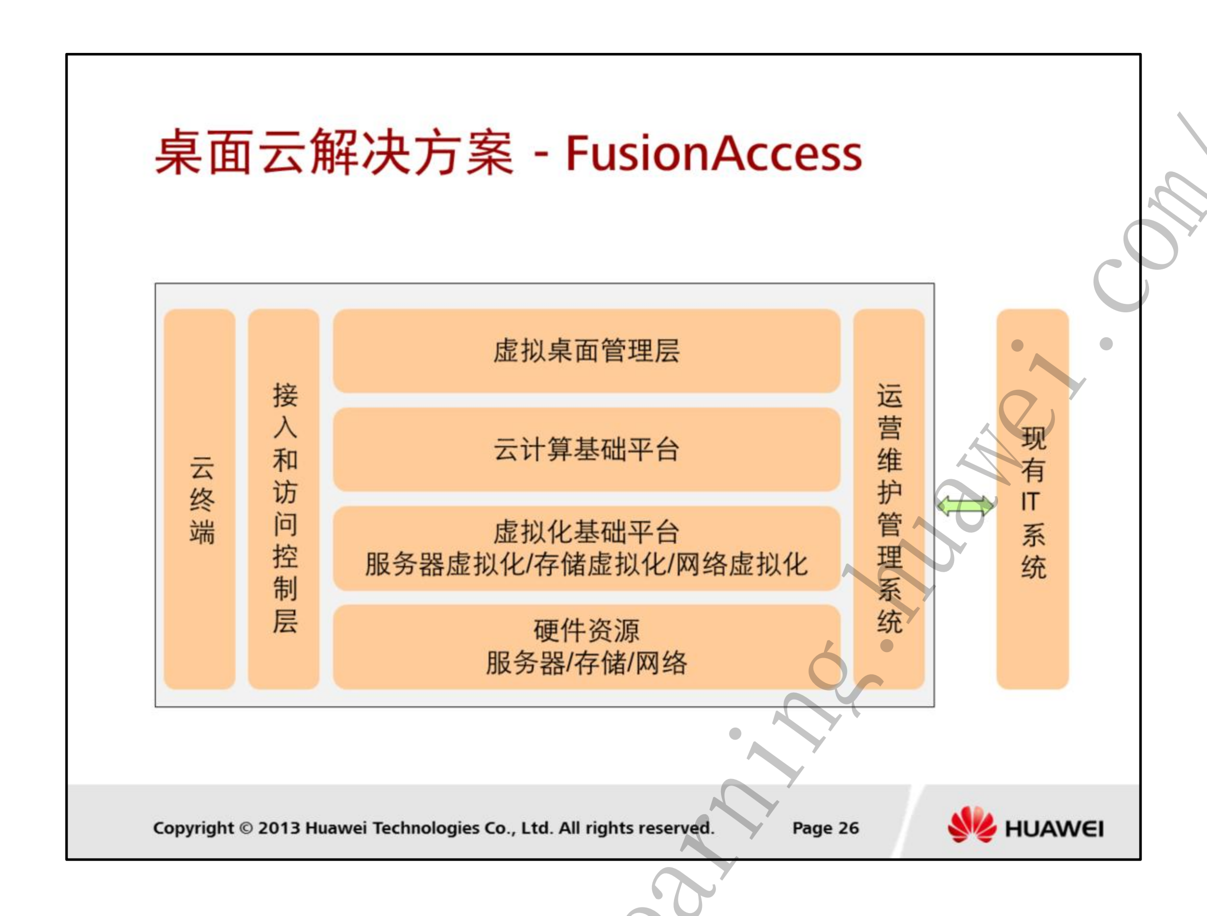
- 虚拟化基础平台

根据虚拟桌面对资源的需求，把桌面云中各种物理资源虚拟化成多种虚拟资源的过程。

- 云计算基础平台

云计算基础平台包含资源管理和资源调度两部分：

  - 云资源管理指桌面云系统对用户虚拟桌面资源的管理。可管理的资源包括计算、存储和网络资源等。
  - 云资源调度指桌面云系统根据运行情况，将虚拟桌面从高负载物理资源迁移到低负载物理资源的过程。



- 虚拟桌面管理层

负责对虚拟桌面使用者的权限进行认证，保证虚拟桌面的使用安全，并对系统中所有虚拟桌面的会话进行管理。

- 接入和访问控制层

用于对终端的接入访问进行有效控制，包括接入网关、防火墙、负载均衡器等设备。

- 运营维管系统

运营维管系统包含业务运营管理和OM管理两部分：

  - 业务运营管理完成桌面云的开户、销户等业务发放过程。
  - OM管理完成对桌面云系统各种资源的操作维护功能。

- 云终端

用于访问虚拟桌面的特定的终端设备，包括瘦客户端、软终端、移动终端等。



总结

- 云计算的概念
- 云计算的部署模式与商业模式
- 云计算的关键技术
- 华为云计算解决方案





思考题

1. 网盘属于云计算商业模式中的哪一种?
2. 结合华为FusionSphere解决方案，谈谈云计算的价值有哪些?



练习题

- 多选题

  1、云计算的部署模式有哪几种？（ ）

  A、私有云

  B、公有云

  C、混合云

  D、桌面云

  2、云计算的商业模式有哪几种？（ ）

  A、IaaS

  B、PaaS

  C、SaaS

  D、DaaS



答案：

1、ABC

2、ABCD



练习题

判断题

1、 华为数据中心属于公有云（　　）

2、华为FusionCube解决方案集计算、存储、网络于一体。（ ）



答案：

1、F

2、T

Thank you

www.huawei.com

华为职业认证通过者权益

通过**任一项**华为职业认证，您即可在华为在线学习网站(*http://learning.huawei.com/cn*) 享有如下特权：

- 1、**华为E-learning 课程学习**
  - **内容**：**所有华为职业认证E-Learning课程**，扩展您在其他技术领域的技术知识
  - **方式**：请提交您的“华为账号”和注册账号的“email地址”到 *Learning@huawei.com 申请权限。*
- 2、**华为培训教材下载**
  - **内容：华为职业认证培训教材+华为产品技术培训教材，**覆盖企业网络、存储、安全等诸多领域
  - **方式**：登录*华为在线学习网站*，进入“*华为培训->面授培训*”，在具体课程页面即可下载教材。
- 3、**华为在线公开课(LVC)优先参与**
  - **内容**：企业网络、UC&C、安全、存储等诸多领域的职业认证课程，华为讲师授课，开班人数有限
  - **方式**：开班计划及参与方式请详见LVC排期：
    http://support.huawei.com/learning/NavigationAction!createNavi#navi[id]=_16
- 4、**学习工具 eNSP**
  - *eNSP (Enterprise Network Simulation Platform)*, 是由华为提供的免费的、可扩展的、图形化网络仿真工具。主要对企业网路由器和交换机进行硬件模拟，完美呈现真实设备实景；同时也支持大型网络模拟，让大家在没有真实设备的情况下也能够进行实验测试。
- 另外，华为建立了知识分享平台 华为认证论坛。您可以在线与华为技术专家交流技术，与其他考生分享考试经验，一起学习华为产品技术。（http://support.huawei.com/ecommunity/bbs/list_2247.html ）

